UNIVERSIDADE DA BEIRA INTERIOR
Covilhã | Portugal

Escul@pio:Uma plataforma Colaborativa de Acesso ao UMLP

Ruben Costa

Submitted to University of Beira Interior in candidature for the degree of Master of Science in Informatics Engineering

Supervised by PhD Gaël Dias

Departamento de Informática
University of Beira Interior
Covilhã, Portugal
http://www.di.ubi.pt

Agradecimentos

À Universidade da Beira Interior e ao departamento de Informática por me permitir aprender e crescer durante estes últimos anos de formação, qualificando-me para realizar este trabalho.

Ao meu orientador Prof. Dr. Gaël Dias e a Isabel Marcelino (PhD) pela orientação dada, assim como a todo o pessoal do Hultig por estar sempre pronto a dar uma ajuda quando necessária.

A minha família e aos meus amigos por todo o apoio

Resumo

O UMLP surge com a ideia de acabar com os problemas de comunicação na sociedade médica, criando um léxico unificado de termos médicos. Os termos são extraídos de fontes cuja fidelidade seja garantida (Wikipédia, Wikcionário, Priberam, Médicos de Portugal, DeCS, Glossário Medico). Em particular são criados sistemas de extracção de informação para cada uma das fontes de informação. Uma vez extraídos os termos uma vez extraídos são analisados, corrigidos e é criado um léxico unificado.

Uma vez o léxico unificado é preciso criar plataformas capazes de levar até ao utilizador o acesso à informação, plataformas essas que têm que ser simples, práticas, intuitivas e visualmente agradáveis. É então criado o Escul@pio, uma plataforma colaborativa de acesso ao UMLP, também disponivel a partir de dispositivos móveis

Conteúdo

Lista de Figuras

Lista de Tabelas

Acrónimos

API - Application Programming Interface

BIREME - Biblioteca Regional de Medicina

CKB - Collaborative Knowledge Bases

Decs - Descritor em Ciências da Saúde

DPLP - Dicionário Priberam da Língua Portuguesa

DTD - Document Type Definition

FLiP - Ferramentas para a Língua Portuguesa

HTML - HyperText Markup Language

ICD-10 - The International Statistical Classification of Diseases and Related Health Problems 10th Revision

ICF - International Coach Federation

INESC-ID - Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento

JWPL -Java Wikipedia Library

JWKTL -Java based Wiktionary Library

LILACS - Literatura Latino-Americana e do Caribe em Ciências da Saúde

LKB - Linguistic Knowledge Bases

MEDLINE - Medical Literature Analysis and Retrieval System Online

MeSH - Medical Subject Heading

NLM - National Library of Medicine

NLP - Natural Language Processing

PDA - Personal Digital Assistants

PFIBF - Path Frequency - Inversed Backward Link Frequency

UMLF - Unified Medical Lexicon for French

UMLP - Unified Medical Lexicon for Portuguese

UMLS - Unified Medical Language System

UMLSKS - Unified Medical Language System Knowledge Source Server

RIA - Rich Intercative Applications

SIGWP -Special interest Group on Wikipédia Research

WWW - World Wide Web

XML - eXtensible Markup Language

Capítulo 1

Introdução

"Information is a source of learning. But unless it is organized, processed, and available to the right people in a format for decision making, it is a burden, not a benefit"

William Pollard

1.1 Problemática

As novas tecnologias, como as redes de alta velocidade e as grandes capacidades de armazenamento a baixo custo, combinado com a grande expansão da Internet, proporcionam um o nível importante de informação. O mesmo se passa com a informação médica, que está disponível através de várias fontes. No entanto, a informação só tem importância caso seja acessível, e seja do interesse do utilizador.

O *Webster's Third International Dictionary* consiste em aproximadamente 500 000 entradas, dentre das quais 200 000 podem ser consideradas como pertencentes ao domínio da linguagem técnica médica [2]. A acrescentar a isto, a utilização de expressões compostas em vez de palavras simples é muito comun na linguagem médica. É seguro dizer que as áreas da saúde sofrem do excesso de informação, em que o número e a diversidade de fontes de informação são muito grandes, originando assim um vasto e complexo léxico, provocando inexoravelmente ambiguidades lexicais.

Numa era em que o recurso à informática assume um papel cada vez mais importante e indispensável no processo clínico, para o qual a margem de erro é muito reduzida, é preciso encontrar métodos para facilitar o acesso e compreensão da informação. Superar certos obstáculos lexicais é muito importante, pois a ambiguidade de termos,

ou de definições de um termo, é um grave problema. Vários termos podem ter a mesma definição ou um termo ter duas ou mais definições distintas, dependendo do seu contexto. É portanto necessário definir e unificar um léxico para que muitos problemas de comunicação se resolvam. Quando a margem de erro é praticamente inexistente, a comunicação é muito importante e é necessário que todas as pessoas envolvidas compreendem e saibam como comunicar.

Assim qualquer pessoa, desde um profissional da saúde, um estudante de medicina ou até pessoas que não estão ligadas directamente ao ramo da saúde, precisam ter livre acesso à informação de maneira simples e compreensível, para que não existam equívocos na comunicação.

1.2 Objectivo

Este trabalho tem como objectivo principla a construção de uma plataforma colaborativa de acesso ao léxico unificado do português, o (UMLP- Unified Medical Lexicon for Portugues). Trata-se de um dicionário de termos médicos com a finalidade de facilitar o acesso à informação e eliminar problemas de ambiguidade lexical. Este projecto é composto por três etapas: recolha, análise e publicação da informação como podemos ver na figura 1.1.

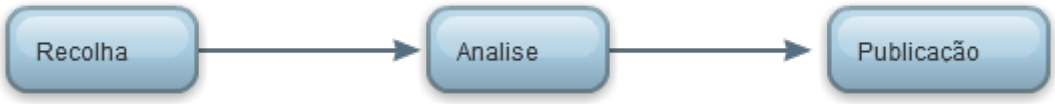

Figura 1.1: Etapas da construção de todo o projecto

Como dito anteriormente a informação só tem importância se esta for acessível ao utilizador e de fácil compreensão. Por isso, recorrendo às mais recentes tecnologias, foi criado o "Escul@pio". O Esculápio é o deus da medicina e da cura na mitologia greco-romana, aqui adoptado como o nome para a nossa aplicação. Esta plataforma online é um dicionário colaborativo, que permite inserir informação sobre um termo ou alterar uma definição existente sendo esta decisão partilhada com todos os membros da rede.

Uma aplicação para dispositivos móveis foi também estudada de modo a que em qualquer lugar o utilizador possa aceder à base de dados de termos médicos unifi-

cados, concedendo assim uma maior mobilidade na pesquisa da informação. Pois os profissionais de saúde, devido à natureza do seu trabalho, necessitam de uma grande mobilidade, e é um facto que os estudantes de medicina estão rapidamente adoptando o uso de PDA (Personal Digital Assistants) para aceder à uma variedade de informações [3], leva a que este trabalho vá de encontro as tendências dos seus utilizadores, sempre com o objectivo de facilitar o acesso a informação.

1.3 Metodologia

Este projecto começa com a recolha de termos de fontes cuja integridade da sua informação seja garantida, para que não existam incorrecções ou incoerências. As fontes usadas são agrupadas em dois grupos. Colaborativa, criadas por voluntários, várias pessoas que partilham o conhecimento acrescentado termos, corrigindo e adicionado definições (Wikipedia[1] , Wikcionário[2]), e não colaborativos que são fontes de informação cujo léxico é criado por um número limitado de profissionais linguistas (Priberam[3], Decs[4], Médicos de Portugal[5], Glossário Multilingue de Termos Médicos Técnicos e Populares[6]) como podemos ver na figura 1.2.

Após a recolha de todo o vocabulário, é procedido à análise e correcção ortográfica da mesma, remoção de alguns termos fora do domínio da saúde. Pois devido ao facto de haver tantas fontes de informação é normal que algumas ambiguidades e inconsistências apareçam. Por fim, é feita uma interligação entre os termos das diversas fontes, chamada de unificação do léxico. Toda esta etapa do projecto é feita manualmente, e enquadra-se na tese de Doutoramento da estudante Isabel Marcelino.

Uma vez recolhida e unificada toda a informação, estamos na posse de uma léxico de grande rigor científico, e com termos definidos de forma mais completa e a respectiva conexão com os seus sinónimos, antónimos, etimologias e classificação taxonómica.

[1] www.wikipedia.org
[2] www.wiktionary.org
[3] www.priberam.pt
[4] www.priberam.pt
[5] www.Decs.bvs.br
[6] http://users.ugent.be/ rvdstich/eugloss/PO/lijsta.html

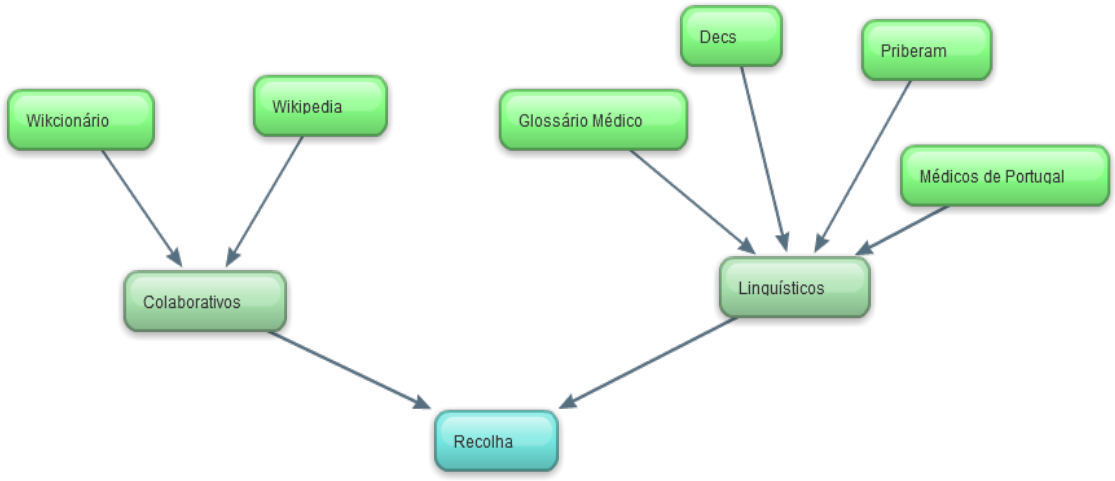

Figura 1.2: Agrupar as diversas fontes de informação em dois grupos

1.4 Plano da Tese

A tese está organizada da seguinte forma. No capítulo 2, é feita uma descrição do trabalho que já existe nesta área, são apresentados os problemas dessas aplicações e é apresentada uma proposta de solução para os problemas encontrados. No capítulo 3, são é apresentadas as abordagens, problemas e soluções encontradas para a extracção da informação das várias fontes de informação, são também apresentadas as diferenças entre os tipos de fontes. Nos capítulos 4 e 5, são apresentados os aspectos da interface tanto PC como do PDA, como alguns exemplos de dicionários existentes, e os aspectos mais técnicos da arquitectura. No capítulo 6, é feita a conclusão de todo o projecto e a apresentação de ideias que ficam por implementar.

Capítulo 2

Trabalho Relacionado

A integração de terminologias padrão num sistema de representação de conhecimento unificado da medicina tem sido uma área chave da investigação nos últimos anos. O *Unified Medical Language System* (UMLS) concebido pela *National Library of Medicine* (NLM) em Bethasda, nos Estados Unidos, é um dos maiores esforços nesse sentido, conseguindo combinar um grande número de termos numa única plataforma. Existem no entanto outros esforços para a recolha e organização de termos do domínio da medicina. O Descritores em Ciências da Saúde (DeCS) criado pela BIREME é outra plataforma que contem um elevado número de termos médicos para o Português.

2.1 Unified Medical Language System

O UMLS é um repositório de termos relacionados com a área biomédica, desenvolvida pela *US National Library of Medicine*. O UMLS é um projecto que teve o seu aparecimento em 1986, na sua forma primitiva. Actualmente o UMLS tem mais de 2,5 milhões de termos para mais de 1 milhão de conceitos em mais de 100 fontes de informação, contendo aproximadamente 12 milhões de relações entre os conceitos [4].

A NLM desenvolveu o UMLS como um esforço para superar duas barreiras significativas: a recuperação de informação legível por máquina (exite uma variedade de nomes usados para expressar o mesmo conceito) e a falta de um formato padrão para a distribuição de terminologias [5].

O UMLS está dividido em três grandes componentes como podemos ver na figura 2.1: o Metathesaurus, a Rede Semântica e o léxico SPECIALIST. As componentes

podem ser usadas em conjunto ou separadamente.

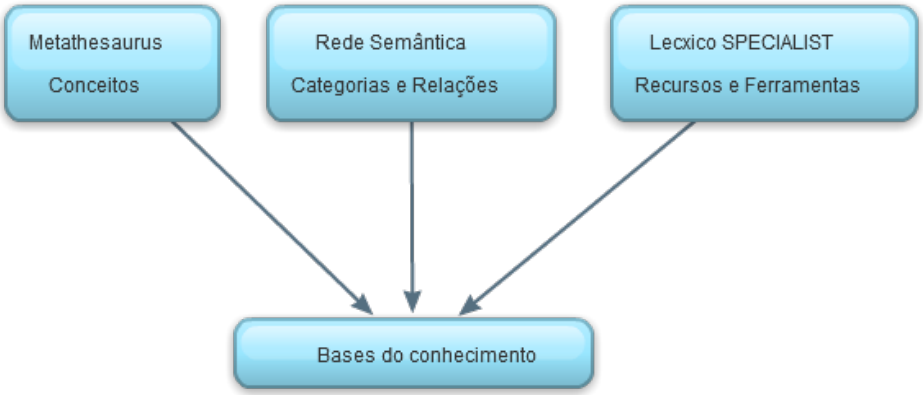

Figura 2.1: As três componentes que constituem o UMLS

2.1.1 Metathesaurus

O Metathesaurus é uma base de dados multilingue que contem informação sobre conceitos médcos, as suas várias instâncias e o relacionamento entre eles. Este tesauro é constituído a partir de 100 lexicos. A figura 2.2 ilustra como o Metathesaurs integra estas terminologias, pode servir com elo de ligação entre eles e os subdomínios que eles representam [4].

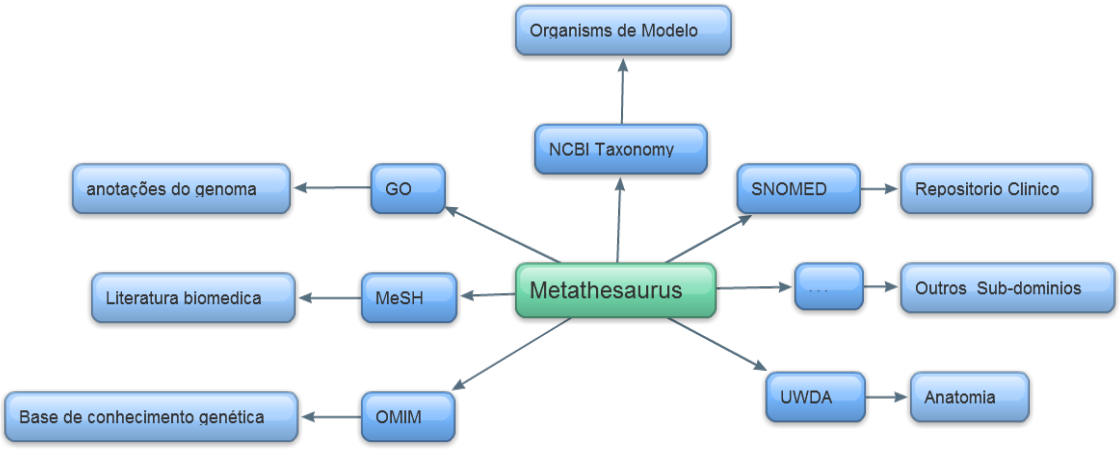

Figura 2.2: Os vários subdomínios que constituem o Metathesaurus

O Metathesaurus unifica diferentes terminologias e organiza-as por conceitos ou significados, criando ligações de nomes alternativos de um mesmo conceito. Também

identifica as relações entre diferentes conceitos. Quando duas fontes de informação utilizam o mesmo conceito, o Metathesaurus representa ambos os significados e indica em que terminologia o significado está presente. Quando o mesmo conceito está presente em diferentes contextos hierárquicos, o Metathesaurus inclui todas as hierarquias do conceito. O Metathesausus não representa uma única visão consistente do mundo, ele preserva os muitos pontos de vista das diferentes fontes de informação, porque estes podem ser úteis para diferentes tarefas[1].

Um significado pode ter diferentes instâncias. O objectivo do tesauro é entender o significado de cada instância em cada fonte e ligar essas que significam o mesmo. Quando é feita uma pesquisa no UMLS a partir da ferramenta disponível na internet *Unified Medical Language System Knowledge Source Server*(UMLSKS), vão aparecer não só os vários conceitos como também os sinónimos encontrados pelo tesauro[2].

2.1.2 Rede Semântica

A Rede Semântica consiste num vasto conjunto de tipos semânticos, que proporcionam uma categorização consistente de todos os conceitos representados no Metathesaurus, e promove os relacionamentos entre estes conceitos. O âmbito da Rede Semântica é amplo, permitindo a categorização semântica de uma vasta gama de terminologias em vários domínios, fornecendo informação sobre os tipos semânticos que podem ser associados aos conceitos e define um conjunto de relacionamentos entre os tipos semânticos. Esta rede contem 133 tipos semânticos e 54 relacionamentos[3]. A Rede Semântica serve de autoridade para os tipos semânticos que são atribuídos aos conceitos no Metathesaurus. A Rede define esses tipos, quer seja por descrições textuais ou por informações extraídas da própria hierarquia [6].

Os nós na Rede Semântica são representados pelos tipos semânticos, e as ligações existentes entre os nós fornecem os tipos de relacionamento existentes na Rede. A categorização semântica do UMLS é muito ampla, e cobre múltiplos domínios de terminologia como por exemplo, organismos, estruturas anatómicas, funções biológicas, químicas, eventos clínicos, objectos físicos, conceitos e ideias.

Como visto em 2.1.1 o Metathesaurus é constituído por termos oriundos de 100

[1] http://www.nlm.nih.gov/research/umls/umlsdoc.htm
[2] http://www.ncbi.nlm.nih.gov/bookshelf
[3] http://www.nlm.nih.gov/research/umls/umlsdoc.htm

fontes de vocabulário diferentes. O significado destes termos é definido pela sua fonte, explicitamente por definição ou anotações (por contexto i.e. o seu lugar no tesauro, por sinónimos ou palavras relacionadas). A cada conceito do Metathesaurus é também atribuído um tipo semântico, o tipo semântico mais especifico disponível na hierarquia. Por exemplo, o conceito "Macaco" recebe o tipo semântico "Mamífero", porque não existe um tipo específico como "Primata", disponível na Rede [4].

2.1.3 Léxico SPECIALIST

A terceira componente que constitui o UMLS é o léxico SPECIALIST que é um vocabulário em inglês composto por palavras seleccionadas de uma variedade de fontes: 20.000 palavras do *UMLS Test Collection of MEDLINE absctacts*, juntamente com as palavras que aparecem no Metathesaurus e no *Dorland's Illustrated Medical Dictionary*. É também composto por palavras do vocabulário geral, as 10.000 mais frequentes do *The American Heritage Word Frequency Book* e a lista das 2.000 palavras usadas nas definições do *Longman's Dictionary of Contemporary English*[5] [7].

O léxico regista para cada palavra ou expressão a informação sintáctica, morfológica e ortográfica necessária para um futuro tratamento linguístico. Os elementos lexicais podem ser termos simples ou compostos, expansões ou abreviações e siglas.

O léxico SPECIALIST tem como objectivo proporcionar informação lexical necessária para o sistema SPECIALIST NLP (Natural Language Processing). Esta ferramenta foi projectada para lidar com o elevado grau de variabilidade das palavras da linguagem natural. Muitas vezes as palavras têm várias formas flexionadas que são consideradas instâncias da mesma palavra [6]. Por exemplo o verbo *treat*, tem três variantes flexionais:

- *Treats* - a terceira pessoa do singular do presente
- *Treated* - a forma do passado e particípio passado
- *Treating* - a forma de particípio presente

Os termos multi-palavras do Metathesaurus podem também ter várias ordens das palavras, além de variantes em casos flexionais e alfabéticos. As ferramentas lexicais permitem ao utilizador abstrair-se destas variações.

[4] http://www.ncbi.nlm.nih.gov/bookshelf
[5] http://www.nlm.nih.gov/research/umls/umlsdoc.htm
[6] http://www.nlm.nih.gov/research/umls/aboutumls.htm

2.2 UMLF - Unified Medical Lexicon for French

Recursos básicos da linguagem natural como os do UMLS, são elementos chaves para a informática virada para a medicina. Para além do UMLS construido para o inglês, já foi iniciada a construção de uma versão em alemão [8], e outro está planeado para espanhol [9]. Para a língua francesa existem alguns recursos, mas estão incompletos e espalhados por vários domínios. O UMLF (Unified Medical Lexicon for French) fundado pelo Ministério Francês da Investigação e Educação, tem como objectivo reunir e unificar vários recursos, completá-los, e torná-los disponíveis num formato standard, para a indústria e investigadores [10]. Para a construção do léxico, a linguagem médica foi recolhida por meio de análise de grandes corpora diversificados, representando diversas especialidades médicas, e através da compilação de lexicos medicos controlados, como por exemplo ICD-10(*The International Statistical Classification of Diseases and Related Health Problems 10th Revision*), ICF (*International Coach Federation*), SNOMED Francês, o Catalogo de procedimentos (CCAM), VIDAL thesauru (VidalCIM) com também o recentemente criado MeSH Francês. As palavras no léxico são palavras isoladas, mas também palavras compostas. O léxico vai conter para cada palavra com informação gramatical (substantivo, adjectivo, género, numero, etc) [11].

O objectivo é providencias o acesso as principais terminologias medicas em francês, juntamente com métodos de indexação controlada.

2.3 DeCS - Descritores em Ciência da Saúde

O DeCS (Descritor em Ciências da Saúde) foi criado pela BIREME, Centro Latino-Americano e do Caribe e de Informação em Ciências da Saúde, para uso na indexação de artigos de revistas científicas, livros, anais de congressos, relatórios técnicos, e outros tipos de materiais, assim como para ser usado na pesquisa e recuperação de assuntos da literatura científica nas bases de dados: LILACS (*Literatura Latino-Americana e do Caribe em Ciências da Saúde*), uma base de dados que abrange toda a literatura relativa as ciências da saúde, produzida por autores latino-americanos; MEDLINE (*Medical Literature Analysis and Retrieval System Online*), é a base de dados bibliográficos da NLM; e outras.

Foi desenvolvido a partir do MeSH (Medical Subject Headings da U.S. National Library of Medicine) com o objectivo de permitir o uso de uma terminologia comum para

pesquisa em três idiomas, inglês, espanhol e português (BR), proporcionando um meio consistente e único para a recuperação da informação independentemente do idioma [12].

A primeira versão do DeCS é datada de 1987, no formato impresso, apresentada em dois volumes constituídos pelas listas alfabéticas e hierárquicas, nos idiomas português (Br) e espanhol. A partir do ano 1999, o DeCS, formado pelas listas alfabéticas permutada e hierárquica, foi disponibilizado na versão online [13].

Podemos considerar que o DeCS é um lexico traduzido do MeSH para o português e espanhol, é uma linguagem poli-hierárquica que possibilita a um mesmo descritor estar agrupado em mais do que uma categoria. O DeCS é um vocabulário em crescimento, e actualmente conta com cerca de 30 369 descritores, sendo destes 25 671 retirados do MeSH e 4 698 são exclusivamente do DeCS [12].

O DeCS é uma das fontes usadas neste projecto uma vez que é um vocabulário de termos médicos em português (BR). É discutido mais em detalhe no próximo capítulo.

2.4 Problemas

Todos os trabalhos apresentados neste capítulo, são trabalhos de grande rigor científico, feitos à mão, o que requer muitos recursos, principalmente humanos, precisando de pessoas especializadas para as tarefas de criação do léxico e de o manter actualizado. O que leva a que sejam projectos que embora tenham um controlo de qualidade muito elevado são também muito dispendiosos.

A actualização também não é um processo instantâneo, uma vez que muitos léxicos só são actualizados uma vez por ano na melhor das hipóteses, o que leva a um grande nível de desactualização em certos casos.

2.5 Solução

Em português assim como no caso do francês como discutido anteriormente, existe alguns recursos, mas eles são muito diversos e separados por vários domínios. Para criar um léxico médico é preciso encontrar e unificar os recursos existentes. Criando um sistema de recolha automática, em certos domínios de qualidade certificada, é possível criar um vasto léxico garantindo o rigor científico, e dispensando o trabalho manual

que é demorado e dispendioso.

Para que esta seja uma solução sempre actual, a possibilidade de, a qualquer altura, procurar nas fontes originais actualizações, é uma possibilidade. Também a possibilidade de o utilizador do léxico o poder alterar, permite assim obter um léxico actualizado e rigoroso.

Capítulo 3

Extracção da Informação

Para a elaboração deste trabalho foram usadas várias fontes de informação, com o objectivo de no final ter um léxico unificado de grande rigor científico e de elevado nível de grandeza.

Para que exista um léxico médico, o primeiro passo é a procura e extracção de informação relevante que esteja dentro do domínio da linguagem médica. Como a WWW (*World Wide Web*) é muito vasta, diversa e muito dinâmica, o que por vezes torna-se um problema encontrar informação relevante. Para isso foram criados diferentes *crowlers*, com a capacidade de procurar e extrair informação relevante, criando assim uma nova base de conhecimento através da informação disponível na Web [14].

3.1 Especificidades das Fontes de Informação

A informação usada para a criação do léxico unificado em português é proveniente de várias fontes cuja integridade da sua informação é garantida. Estas fontes podem ser classificadas em dois grupos: base de conhecimento colaborativas (*Collaborative Knowledge Bases - CKB*) e bases de conhecimentos linguísticos (*Linguistic Knowledge Bases - LKB*) como referido em [15][1]. As propriedades das CKB são diferentes das LKB em vários aspectos. A tabela 3.1 mostra uma visão global das características de cada uma delas.

As LKB são tipicamente construídas por linguistas seguindo o mesmo modelo teórico, enquanto que os CKB são construídas por voluntários não profissionais seguindo orientações não vinculativas. Uma abordagem de construção menos rigorosa resulta

Tabela 3.1: Comparação entre CKB e LKB [1]

	Bases de Conhecimentos Linguísticos	Bases de Conhecimento Colaborativas
Construtores	Linguístas	Principalmente não profissionais voluntários
Abordagem de construção	Seguindo modelos teóricos e evidências no corpus	Seguindo orientações não vinculativas
Custos de construção	Significativos	Praticamente inexistentes
Tamanho	Limitado pelos custos de construção	Muito grande e de rápido crescimento
Qualidade da informação	Controlo editorial	Controlo social pela comunidade
Linguagens disponíveis	Línguas principais	Muitas línguas interligadas

em algumas vantagens:

- As CKB são normalmente disponibilizadas segundo licenças que garantem liberdade no seu uso, enquanto as LKB são por norma mais restritas na distribuição devido aos seus custos de construção e manutenção.
- As CKB estão em constante actualização, enquanto que os ciclos de actualização dos LKB não conseguem estar actuais em eventos recentes.
- CKB populares como a Wikipédia ou o Wikcionário são geralmente muito maiores comparando com as LKB.
- CKB estão disponíveis numa grande variedade de línguas interligadas, que os LKB podem não disponibilizar.

No entanto também existem algumas contrapartidas na utilização de CKB em comparação com as LKB:

- As LKB são melhor estruturadas do que as CKB.
- As LKB têm muito menos ruído do que as CKB.

- As CKB estão dependentes do controlo social para manter a precisão e compreensividade da informação, enquanto que as LBK por norma garantem um controlo de qualidade através de profissionais da área.

3.1.1 Bases de Conhecimentos Colaborativas

As bases de conhecimento colaborativas, como dito anteriormente, são construídas através de voluntários, muitas vezes não profissionais na área. O Wikipédia e Wiktcionário que pertencem a *Wikimedia Foundation*, uma organização benéfica sem fins lucrativo, dedicada a incentivar a produção, desenvolvimento e distribuição de conteúdos livres e multilingue [1], são duas das fontes de vocabulário usadas no UMLP.

Tanto o Wikipédia como o Wikcionário são serviços colaborativos que permitem aos voluntários adicionarem, editarem e a apagarem artigos consoante o seu conhecimento referente ao tema do artigo em questão. Um artigo pode ter vários autores que vão editando sucessivamente um mesmo artigo. A *Wikimedia Foundation* tem como lema "Imagine um mundo em que cada ser humano tenha livre acesso à soma de todo o conhecimento".

Devido à sua estrutura colaborativa, o Wikipédia e o Wiktcionário são vítima de vários ataques à confiabilidade da informação que partilham, porque qualquer utilizador pode escrever um artigo e publicá-lo, o que leva a que falsos artigos possam surgir. Sendo eles uma fonte de informação que à partida não oferece garantias de fiabilidade, como podem ser usados como base para um dicionário médico?

É verdade que a estrutura colaborativa facilita o aparecimento do chamado "vandalismo". No entanto, estudos feitos por várias entidades concluíram que o próprio sistema que provoca tanta desconfiança, é também responsável por analisar e corrigir qualquer caso de erros ou vandalismo. Defacto, um colaborador pode assumir vários níveis de colaboração, em actividades tais como: escrever, corrigir falhas e erros ortográficos, traduzir artigos e divulgar ideias ou participar em discussões pertinentes. Assim casos de vandalismo são geralmente corrigidos ou eliminados por um colaborador. Em 2005 um estudo levado a cabo pelo jornal britânico *Nature* [16], mostrou que embora existam erros, o Wikipédia está praticamente ao mesmo nível do que a enciclopédia Britânica. Embora os casos de vandalismo sejam comuns no Wikipédia, por norma estes são corrigidos rapidamente e na maioria dos casos grande parte dos utilizadores nem sofre

[1] http://wikimediafoundation.org

os seus efeitos. O Wikipédia tem uma grande e superintendente capacidade de auto-corrigir-se [17].

3.1.1.1 Wikipédia - Estrutura

O Wikipédia é uma das maiores e mais completas enciclopédias a nível mundial. Foi fundado em 2001 e actualmente possui mais de 15 milhões de artigos em mais de 260 línguas. Em português, possui actualmente mais de 585 000 artigos publicados [2].

O Wikipédia é uma enorme rede de informação. A quantidade de artigos contidos na enciclopédia online é muito grande e esses artigos são de uma grande variedade de temas. Desde o início do Wikipédia, tem havido um esforço para categorizar os seus artigos. O sistema de categorias do Wikipédia está projectado para navegar através de artigos semelhantes. Este sistema de categorização é descrito como uma folksonomia [18], ou seja, um sistema de classificação análogo a uma taxonomia, mas colaborativo, que permite a cada utilizador da informação classificá-la com uma ou mais palavras-chaves, conhecidas como "tags" (em português, marcadores). Este tipo de classificação colaborativa oferece muitas vantagens, não sendo possível no entanto, a uma administração estar responsável pela classificação do conteúdo.

Além disso as categorias, também possuem categorias mais amplas (super-categorias), criando assim uma estrutura hierárquica, a qual se pode chamar de tesauro. Tais relações podem ser adicionadas e removidas pelos utilizadores [19].

A cada categoria pode ser atribuída uma ou mais categorias. Assim sendo o sistema de categorias do Wikipédia não pode ser classificado como árvore, mas sim um grafo direccionado, como vemos na figura 3.1.

Sendo o sistema de categorização do Wikipédia construído com base numa abordagem *bottom-up* [18], cria-se assim um conjunto de vantagens:

- Rápida introdução de novos conceitos: Sem qualquer restrição para a utilização de novas categorias, o número de categorias cresce rapidamente.

- Flexibilidade: Uma vez que o número de categorias por cada artigo do Wikipedia não é limitado, atribuir categorias pode reflectir vários aspectos do conceito.

No entanto, este tipo de categorização também tem as suas desvantagens:

[2] http://www.wikcionary.org

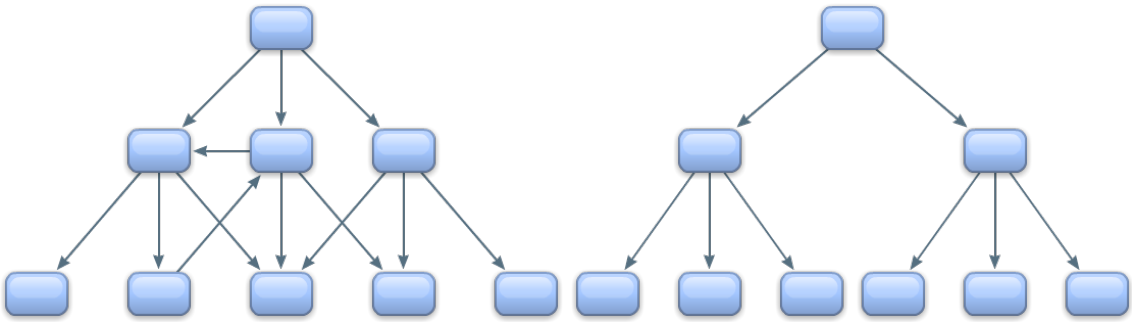

Figura 3.1: Tipos de estruturas de classificação; à esquerda Grafo Direccionado, à direita uma Árvore

- Falta de estabilidade: Uma vez que qualquer pessoa pode editar o Wikipédia, a estrutura de categorias muda rapidamente, então navegar usando a estrutura nem sempre é confiável.

- Fraca estrutura organizacional: Alguns artigos do Wikipédia não estão bem organizados devido ao facto de não existirem categorias apropriadas para eles.

O sistema de categorias do Wikipédia é um tesauro que é desenvolvido colaborativamente e usado para indexar os seus artigos.

3.1.1.2 Extracção da Informação

Para os objectivos do trabalho proposto é necessário a extracção dos conceitos médicos contidos da enciclopédia online. Para isso, é necessário ultrapassar alguns problemas que surgem devido ao elevado número de artigos e a sua estrutura complexa. A figura 3.2[3] mostra uma visualização gráfica da densa estrutura do Wikipédia (categorias e as suas interligações, até ao terceiro grau de profundidade) que ajuda a perceber a complexidade do grafo direccionado do Wikipédia. No centro do grafo está o nó pai, neste caso a categoria Medicina. As páginas ligadas directamente ao nó pai são colocadas no meio da esfera e as páginas ligadas a estas são posicionadas na fronteira da esfera. As ligações são codificadas a cores para representar a profundidade do nó pai.

Um outro modo de visualização é através do trabalho elaborado pela SIGWP(*Special interest Group on Wikipedia Research*). Este grupo criou uma aplicação em Silverlgight

[3]www.chrisharrison.net/projects/clusterball

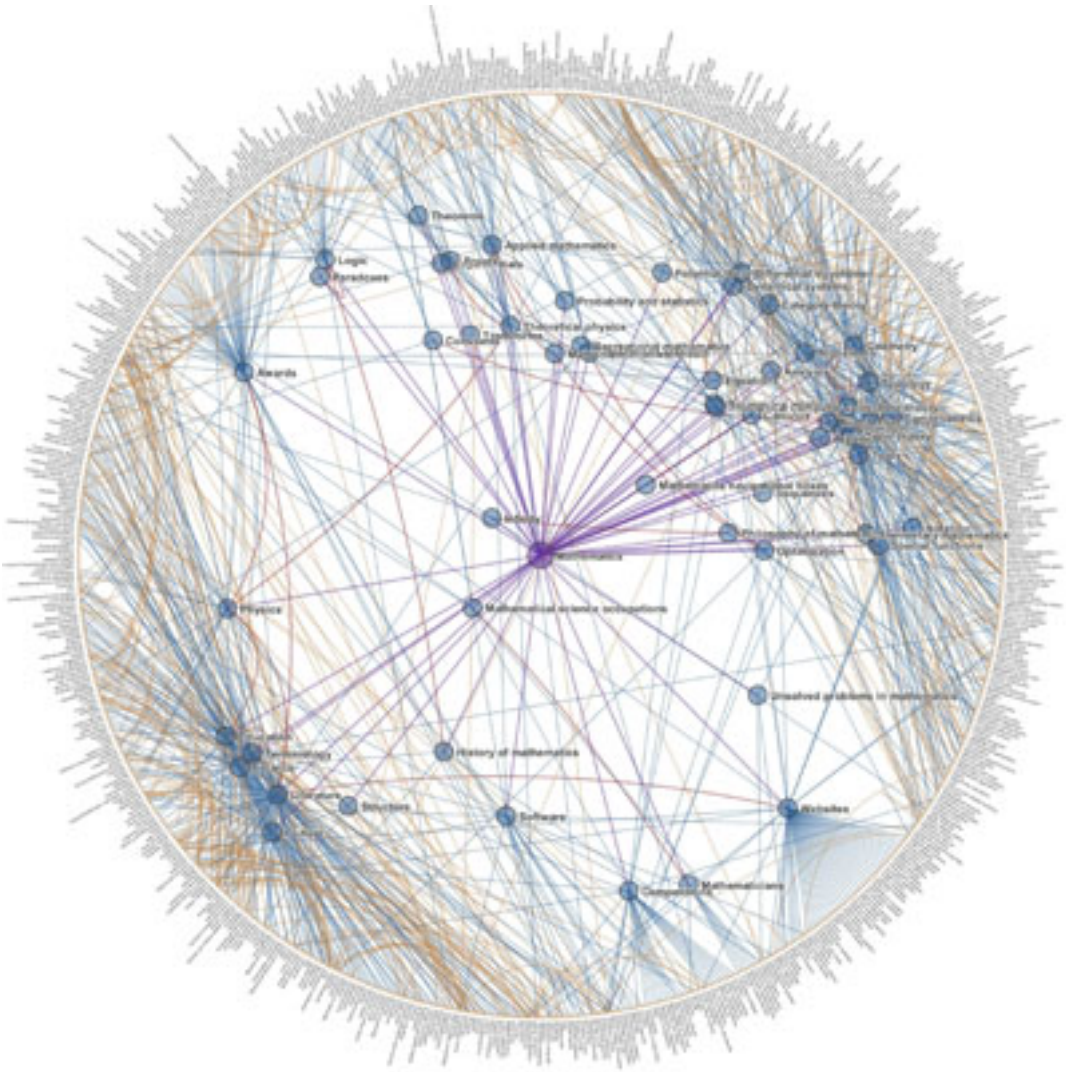

Figura 3.2: ClusterBall Uma representação gráfica da estrutura em categorias do Wikipédia com três níveis de profundidade. No centro encontra-se o nó pai, Medicina.

chamada *Wikipedia Thesaurus Visualizer* [4] que dada um categoria, vai construindo o tesauro que existe a partir dessa categoria. Apenas existe para as versões inglês e japonês do Wikipédia, mas mesmo assim é uma aplicação interessante, ajudando o utilizador a visualizar a rede de relações entre os conceitos.

O nosso objectivo é a partir da categoria Medicina do Wikipédia[5] extrair todos os termos que se encontram nessa categoria e em todas as sub-categorias. Nos artigos do Wikipédia, há uma variedade de informação que é possível extrair para cada termo.

[4] http://wikipedia-lab.org

[5] pt.wikipedia.org/wiki/Categoria:Medicina

Para este trabalho foi recolhido apenas, o termo, a sua definição, url, imagem e a sua legenda caso exista, o caminho partir da origem (categoria Medicina) até ao artigo, a última data de actualização do artigo e o termo nas línguas inglesa, francesa e espanhola. É também guardado um registo da data em que o termo foi extraído do Wikipédia.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE wikipedia SYSTEM "wikipedia.dtd">
<wikipedia>
  <entry id="1">
    <word>Medicina</word>
    <source>wikipedia</source>
    <url doc_date="2 de dezembro de 2009." search_date="5 de Dezembro de 2009" type="html"
      >http://pt.wikipedia.org/wiki/Medicina</url>
    <paths>
      <path>Medicina</path>
    </paths>
    <definition>A medicina é uma das áreas do conhecimento humano ligada à manutenção e restauração
      da saúde. Ela trabalha, num sentido amplo, com a prevenção e cura das doenças humanas num
      contexto médico. É a área de atuação do profissional formado em uma Faculdade de Medicina.
      Segundo a Organização Mundial da Saúde, saúde não é apenas não ter problemas como o cancro,
      doenças crónicas, sida etc. Consiste no bem-estar físico, mental, psicológico e social da
      pessoa. É um estado cumulativo, que deve ser promovido durante toda a vida, de maneira a
      assegurar-se de que seus benefícios sejam integralmente desfrutados em dias posteriores. Nesse
      contexto, diretrizes de organizações supra-nacionais compostas por eminentes intelectuais do
      globo relacionados à área de saúde estabeleceram um novo paradigma de abordagem em medicina. O
      Santo Patrono da Medicina é o Apóstolo São Lucas.</definition>
    <image>
      <legend>O bordão de Esculápio ou caduceu de Asclépio é o símbolo da medicina.</legend>
      <url>http://upload.wikimedia.org/wikipedia/commons/thumb/5/55/Rod_of_Asclepius.svg/180px-Rod_of_Asclepius.svg.png</url>
    </image>
    <categorias>Medicina |</categorias>
    <translation lang="en">
      <word>Medicine</word>
    </translation>
    <translation lang="fr">
      <word>Médecine</word>
    </translation>
    <translation lang="sp">
      <word>Medicina</word>
    </translation>
  </entry>
```

Figura 3.3: Exemplo do ficheiro XML

Existem algumas ferramentas que auxiliam a extracção de informação do Wikiéedia, o JWPL (*Java Wikipedia Library*) é uma dessas ferramentas. É uma API que suporta uma ampla gama de caminhos de acesso, incluíndo interacção entre os artigos, e um eficiente acesso à informação como por exemplo hiperligações e categorias. O JWPL retira a informação directamente da base de dados do Wikipédia, e foi desenhado investigadores em NLP [1].

Embora o JWPL seja uma ferramenta muito poderosa, não foi usada pois não se encontra enquadrada com as necessidades que o projecto tem. Apenas faz a extracção em inglês, e não se limita a um domínio.

Sabendo a informação que é preciso guardar e onde começar a procura do termo, basta desenvolver um *crawler* para este efeito. Alguns motores de busca usam progra-

mas deste tipo para percorrer toda a internet, usando vários em paralelo para conferir maior rapidez [6]. O *crawler* usado não vai percorrer toda a internet, apenas o site do Wikipédia, mais precisamente a partir da categoria Medicina. Dado o url da categoria principal o este vai guardar a informação pretendida e percorrer todas as hiperligações encontradas recursivamente, até não haver mais hiperligações para percorrer.

No entanto existe aqui um grande problema que a princípio passa despercebido. O facto do Wikipédia não estar estruturado como um sistema de classificaçã em árvore controlada, mas sim num grafo social, como mostrado anteriormente, provoca que numa extracção automática da informação, seja possível sair facilmente do domínio da Medicina. Isso obriga a que seja necessário verificar se a categoria actual, pertence ou se está ou não directamente relacionado com o domínio da Medicina. A questão é saber como determinar a cobertura desejada e que informação lexical é útil neste contexto. Embora algumas palavras são nitidamente do domínio médico, outras palavras são muito usadas na linguagem médica, mas não podem ser consideradas especificamente do domínio médico. Por exemplo as palavras "coração", "diagnostico", "cirurgia" e "clínico" são nitidamente do dominio da médico, no entanto, as palavras "direito" e "alargada" são muito usadas no contexto médico mas não são especificas desse domínio.

Para resolver este problema foram estudados alguns algoritmos que encontram relações entre termos no Wikipédia. A WikiRelate [20] é um dos pioneiros no estudo do Wikipedia e a relação entre termos e categorias. Provou que o tamanho do caminho inverso entre termos pode ser usado como medida de relacionamento entre dois dados termos. Porém existem alguns problemas com este algoritmo, nomeadamente à nível de escalabilidade e precisão. A ideia do algoritmo é encontrar o caminho mais curto entre as categorias a que dois conceitos pertencem. No tesauro, como método de medida entre dois conceitos, este método tem resultados interessantes. Contundo, no nosso caso, seria impraticavel procurar neste espaço de dados.

Outro método, um pouco mais complexo do que o anterior é o PFIBF (*Path Frequency-Inversed Backward Link Frequency*) [21]. A ideia do algoritmo é muito simples. A relatividade entre dois artigos v1 e v2 é assumida ser fortemente afectada pelos seguintes factores:

- Número de caminhos do artigo v1 para o artigo v2;
- Comprimento de cada caminho do artigo v1 para o artigo v2;

[6] http://www.webopedia.com/TERM/s/spider.hmtl

A relatividade é forte se houver muitos caminhos entre dois artigos, e se estes forem curtos. Este método tem bons resultados para estabelecer parentesco semântico entre dois termos. No entanto, existe alguma falta de precisão sobretudo quando exista uma forte ambiguidade dos termos. Assim sendo, um forte PFIBF não significa que dois artigos pertençam ao contexto pertendido.

Sendo o Wikipédia uma enciclopédia online com conteúdos muito diversificados, o objectivo é extrair apenas conteúdos ligados unicamente à Medicina. Para isso é preciso criar regras que permitam classificar palavras que são claramente do domínio da Medicina, tendo o cuidado de não omitir outras que sejam, usadas no contexto médico.

Os termos do Wikipédia podem-se dividir em três grupos, os que pertencem nitidamente ao domínio da medicina, os que não são exclusivamente do domínio da medicina, e os que não pertencem ao domínio. Criar um método capaz de analisar um termo e classifica-lo como pertencente a um destes três grupos, é essencial para que o crawler funcione eficientemente.

O método usado é no fundo bastante simples face a complexidade do problema. A ideia é usar o sistema de categorização do Wikipédia para classificar uma página do próprio Wikipédia (artigo ou categoria). Sabendo que cada página do Wikipédia pertence sempre a uma ou mais super categorias, e que estas super categorias indicam o contexto em que uma página se insere. Por exemplo, a categoria Hematológica pertence as super categorias: Especialidades Médicas; Sangue; Biomedicina, que se encontram no final da página como podemos ver na figura 3.3, e são estas super categorias que vão originar a classificação da categoria Hematológia.

O algoritmo para classificar uma página vai analisar as classificações das suas super categorias e assim calcular uma classificação. Para que os algoritmos funcione é preciso resolver dois problemas:

1. Devido ao facto de uma páginas ter mais do que uma super categorias, como achar a classificação da página

2. Como classificar as super categorias.

Analisando o primeiro problema vemos que há varias entradas e é preciso pegar nas entradas todas e transformar numa só classificação, para isso foi construído uma tabela de verdade (tabela 3.2) que visa calcular uma classificação para a página com base em

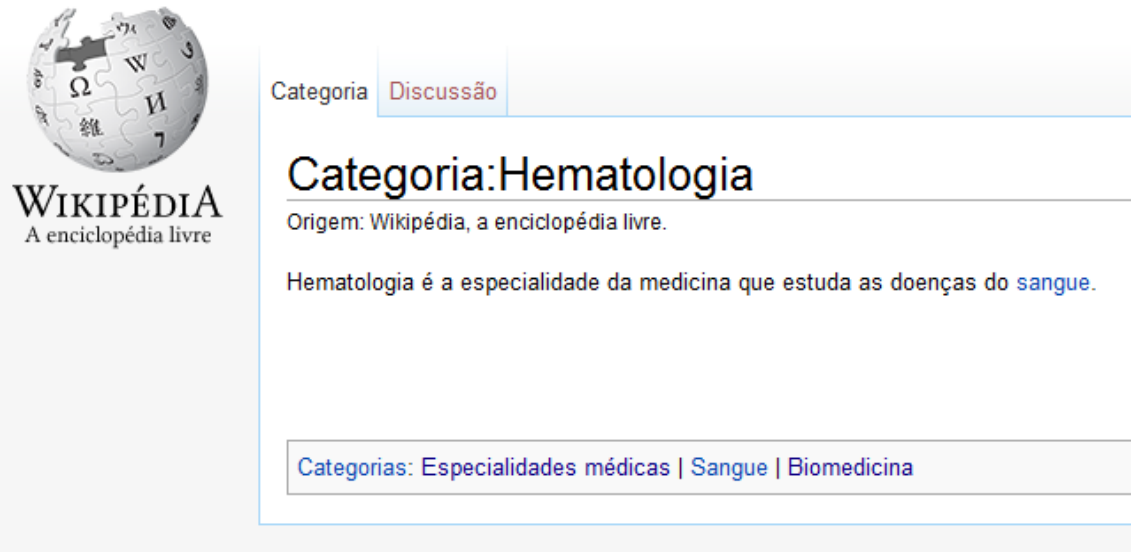

Figura 3.4: Excerto de uma página do Wikipédia.

todas as suas super categorias. A tabela de verdade não usufrui da propriedade de comutatividade e é dada mais importância da esquerda para a direita. A tabela usa um sistema de classificação com cinco níveis classificativos, isto porque o algoritmo tenta procurar de forma abrangente os termos que pertencem ao domínio da medicina, no entanto é preciso um sistema de classificação que vá enfraquecendo a medida que *crawler* vai-se desviando do contexto, no entanto apenas as páginas com a classificação "Não" e "Nunca" são excluídos, a única diferença entre a classificação "Não" e "Nunca" é a força que eles possuem quando o algoritmo calcula o resultado final.

Uma Pagina com, por exemplo, três super categorias cujas classificações são: Sim; Não; Talvez-, a classificação da página vai ser Talvez-, a figura 3.4 mostra como o método de classificação vai calcular o resultado final.

Assim é possível obter um resultado que classifica a página. No entanto, é preciso não esquecer que a ordem pela qual as super categorias são apresentadas é muito importante, pois o resultado pode variar uma vez que consideramos que a primeira super categoria é a que tem mais força dentro do contexto.

O segundo problema prende-se com o facto de arranjar os dados de entrada para que se possa calcular a classificação da página, a ideia é até bastante simples, a medida que o *crawler* vai percorrendo o grafo direccionado recursivamente, e vai classificando as categorias, vai guardando os seus resultados para usar na próxima iteração. Para que seja possível ao *crawler* começar eficientemente, ele começa com

Tabela 3.2: Tabela de verdade para a classificação dos artigos pelas categorias a que pertencem

1-Categoria	2-Categoria	Resultado
Sim	Sim	Sim
	Talvez +	Sim
	Talvez -	Sim
	Não	Talvez +
	Nunca	Talvez -
Talvez +	Sim	Sim
	Talvez +	Sim
	Talvez -	Talvez +
	Não	Talvez -
	Nunca	Não
Talvez-	Sim	Sim
	Talvez +	Talvez +
	Talvez -	Talvez -
	Não	Não
	Nunca	Não
Não	Sim	Talvez +
	Talvez +	Talvez -
	Talvez -	Não
	Não	Nunca
	Nunca	Nunca
Nunca	Sim	Talvez -
	Talvez +	Não
	Talvez -	Não
	Não	Nunca
	Nunca	Nunca

alguns valores iniciais, a categoria "Medicina" começa com a classificação de "Sim", pois é a categoria principal, as suas super categorias (Ciências da saúde, Biologia, Humanos), recebem a classificação de "Talvez +", pois directamente não pertencem ao domínio da medicina, mas os artigos as categorias abaixo têm fortes possibilidades

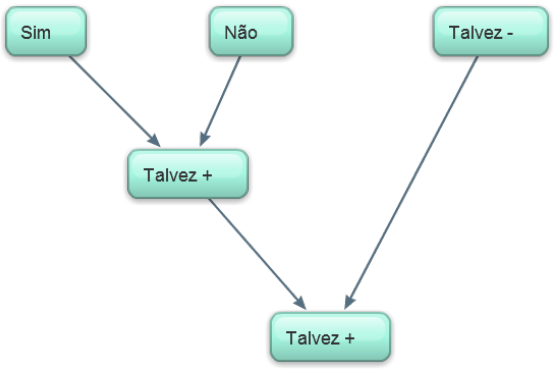

Figura 3.5: Calculo da classificação.

de pertencerem ao domínio, também a categoria "Anatomia" recebe inicialmente a classificação de "Talvez +", também não pertence ao domínio da medicina, pois é muito geral, mas também esta categoria possui muitas categorias abaixo que são claramente do domínio.

Muitas vezes, devido a estrutura do Wikipédia, eventualmente com ciclos ou triângulos, muitas categorias possuem mais do que um caminho desde a raiz. Isso vai provocar que em certos casos o *crawler* vai visitar a mesma página mais que uma vez, nestes casos a classificação que fica registada é a mais elevada.

Podemos ver uma breve explicação do algoritmo no pseudo-codigo.

```
Lista Classificações inicial [Medicina = sim;
 Ciências da saúde = Biologia = Humanos = Anatomia = talvez+]
Abrir página da Wikipédia;
    Calcular classificação;
    Se pertence então
        Adicionar a lista de classificações;
        Extrair informação para o ficheiro XML;
Próxima página;
```

A figura 3.6 mostra um excerto da estrutura do Wikipédia a partir da categoria Medicina e seus filhos, onde é possível ver o resultado do algoritmo de classificação, neste exemplo estão presentes as categorias "Manicure" e "Depilação" cujo domínio

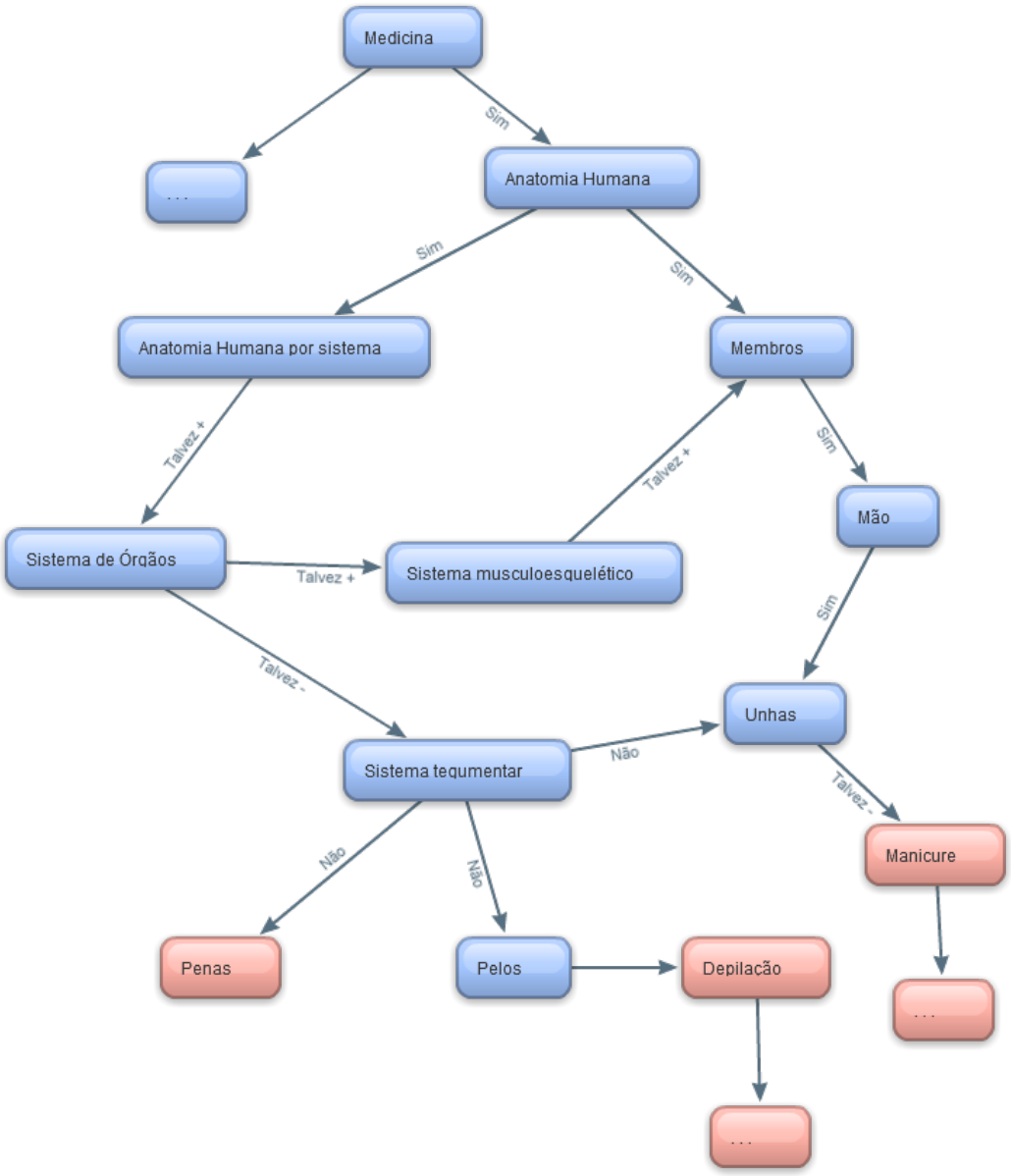

Figura 3.6: Exemplo das categorias do Wikipédia, a relação entre elas e o resultado obtido depois de aplicado o algoritmo de classificação.

não pertence a Medicina, no entanto indirectamente o sistema de categorização do Wikipédia cria uma ligação possível com a categoria Medicina. Como estes, existem muitos outros e mais complexos casos. A figura não mostra as super categorias que constituem cada categoria presente no esquema, portanto a figura não serve para representar o cálculo, uma vez que não mostra os valores de entrada, apenas mostra o resultado. Através da figura podemos ver alguns casos do algoritmo:

1. O caminho entre "Sistema tegumentar" e "Unhas" é classificado como "Não". Isto é porque quando o *crawler* vai do "Sistema tegumentar" para "Unhas", ele ainda não possui a classificação presente nas outras super categorias que constituem a categoria "Unha". Isto acontece porque o *crawler* vai percorrendo o grafo uma categoria de cada vez, e é portanto impossível ter conhecimento da classificação de uma categoria antes de a percorrer, no entanto em casos como a categoria "Unha" que a relação era forte, existe outro caminho, e portanto a sua classificação pode ser actualizada.

2. "Penas" e "Pelos" não são incluídos no domínio. A categoria "Penas" nitidamente é bem excluída do domínio, já a categoria "Pelos" erradamente é excluída.

3. A categoria "Manicure" é incluída no domínio erradamente, no entanto a sua classificação já é muito reduzida, o que leva que as categorias abaixo dificilmente vão ser incluídas no dicionário.

Estes problemas ocorrem essencialmente devido ao facto da categorização do Wikipédia não ser mais específica.

Os artigos que não perecem ao domínio da Medicina e são erradamente incluídos no léxico, são posteriormente excluídos na revisão manual.

O método não é 100% eficaz, como vimos na figura 3.6, e ainda assim é possível encontrar muitos artigos que fujam ao contexto, mas com este algoritmo já é possível eliminar muitas categorias problemáticas que obrigavam ao *crawler* sair do contexto da Medicina e entrar noutro contexto.

Uma vez que o artigo não é rejeitado, então é preciso proceder a extracção da informação. É um processo bastante simples, no entanto é preciso levar em conta alguns pormenores importantes. O *crawler* recupera todo o HTML da página em questão, e vai procurar a informação que é importante para o trabalho, retirando o código HTML e guardando a informação num formato XML. No Anexo A mostra o DTD que define o formato do XML criado para o armazenamento da informação extraída não só do Wikipédia mas também das outras fontes de informação.

OS artigos da Wikipédia, por serem colaborativos, sofrem de algumas diferenças a nível da estrutura HTML, o que dificultou em muito o desenho do *crawler*. Foi preciso levar em conta muitas variações no modo e na ordem como a informação era disposta no HTML da página, para que o *crawler* não fosse levado a extrair informação

incompleta. Contudo isto tudo foi tido em conta para maximizar a quantidade de informação recolhida.

3.1.1.3 Wikcionário - Estrutura

Tal como para o Wikipédia, também existe uma API para extrair informação do Wikcionário, a JWKTL (*Java based Wiktionary Library*). Esta API é em muito semelhante à JWPL. Portanto, tem os mesmos inconvenientes e por isso também aqui não foi usado como ferramenta de extracção [22].

Para o Wikcionário foi usada uma abordagem um pouco diferente da do Wikipédia, devido ao facto do Wikcionário não ter uma estrutura igual a do Wikipédia. Assim como no Wikipédia, também aqui o *crawler* inicia a sua tarefa no início da categoria Medicina[7]. A partir desta página o *crawler*, à semelhança do que foi feito no Wikipédia, vai recolher informação em todos os artigos, passando por todas as subcategorias que encontre, extraíndo todo o HTML de cada pagina e guardando toda a informação disponível para cada termo num formato XML. O que no Wikcionário difere do Wikipédia é que aqui não ficamos só pelos artigos disponíveis dentro das subcategorias. O Wikcionário tem dentro de cada artigo, quando disponível, artigos relacionados. Então o *crawler* vai também dentro de certos artigos, recolher toda a informação disponível. Neste caso, os artigos foram: o próprio artigo de Medicina[8], Remédio[9] e Doenças[10]. Estes artigos foram escolhidos tanto pela sua relevância com o domínio da Medicina com também pela quantidade de artigos a que eles estão relacionados, criando assim um vocabulário muito mais completo e rico.

Ao contraio do Wikipédia aqui não houve o problema do *crawler* sair fora do contexto da pesquisa. Uma vez que o Wikcionário é mais recente, tem menos artigos o que leva a que não exista uma grande profundidade no grafo de relações entre categorias, deixando assim a tarefa do *crawler* facilitada.

O ficheiro XML onde a informação é armazenada possui um DTD igual ao usado no caso do WIkipédia. No entanto, o Wikcionário obriga a umas mudanças na lógica usada até agora. No Wikipédia, cada termo tem uma definição, e cada termo é uma entrada no ficheiro XML. O Wikcionário é diferente. Um termo pode conter várias definições.

[7] http://pt.wiktionary.org/wiki/Categoria:Medicina
[8] http://pt.wiktionary.org/wiki/medicina
[9] http://pt.wiktionary.org/wiki/remédio
[10] http://pt.wiktionary.org/wiki/doenças

Portanto, para cada definição é uma entrada diferente no ficheiro XML. Também, informações como género, número categoria gramatical e etimologia são muito mais comuns de aparecer do que no Wikipédia, uma vez que o Wikcionário é um dicionário que foi desenvolvido com o propósito de ser o complemento lexical do Wikipédia [1].

Finalmente, a semelhança do Wikipédia, o facto de se tratar de um dicionário colaborativo tem como consequência uma possível diferença entre artigos. O *crawler* foi desenhado para tratar estas excepções.

3.1.2 Bases de Conhecimentos Linguísticos

As bases de conhecimentos linguísticos, ao contrário das colaborativas, não estão sujeitas ao chamado "vandalismo", uma vez que os utilizadores não podem modificar o seu conteúdo. A criação do léxico está a cabo de linguistas o que confere ao léxico uma estrutura mais coerente e consistente às bases de conhecimentos colaborativas. No entanto, devido à sua natureza, estes léxicos são rapidamente desactualizados.

3.1.2.1 DeCS - Estrutura

O DeCS, já discutido no capítulo de trabalhos relacionados, foi desenvolvido a partir do MeSH , que é usado para a indexação do corpus MEDLINE no qual são extraídos cerca de 6500 termos, e adiciona mais 5000 termos exclusivos. O DeCS é um léxico estruturado. Os léxicos estruturados são colecções de termos que representam conceitos, organizados segundo uma metodologia na qual é possível especificar as relações entre conceitos com o propósito de facilitar o acesso à informação. Os léxicos estruturados são necessários para descrever, organizar e prover acesso à informação.

O DeCS tal como o MeSH é considerado como um tesauro. A sua estrutura hierárquica é fundamental na divisão do conhecimento em classes e subclasses respeitando conceitos e semânticas.

Para além dos conceitos do léxico MeSH o DeCS adicionou mais quatro categorias, "Ciências da Saúde", "Homeopatia", "Saúde Publica" e "Vigilância Sanitária"[11].

Os conceitos do léxico DeCS estão assim distribuídos (versão 2010):

- 25,8% pertence a compostos químicos e drogas, entendendo aqui tanto as drogas exógenas como as endógenas;

[11] http://decs.bvs.br

- 20,4% pertence à anatomia, organismos, fenómenos e processos;
- 12,9% do total são referentes a doenças;
- 21,6% é a parte das áreas como técnicas e equipamentos, ciências afins, características de publicações e áreas geográficas;
- 18,9% é referentes às categorias adicionadas pelo próprio DeCS i.e. "Saúde Pública", "Homeopatia", "Vigilância Sanitária", "Ciência".

Como podemos ver na figura 3.7[12], a distribuição das várias categorias que compõem o DeCS é a seguinte na sua última actualização.

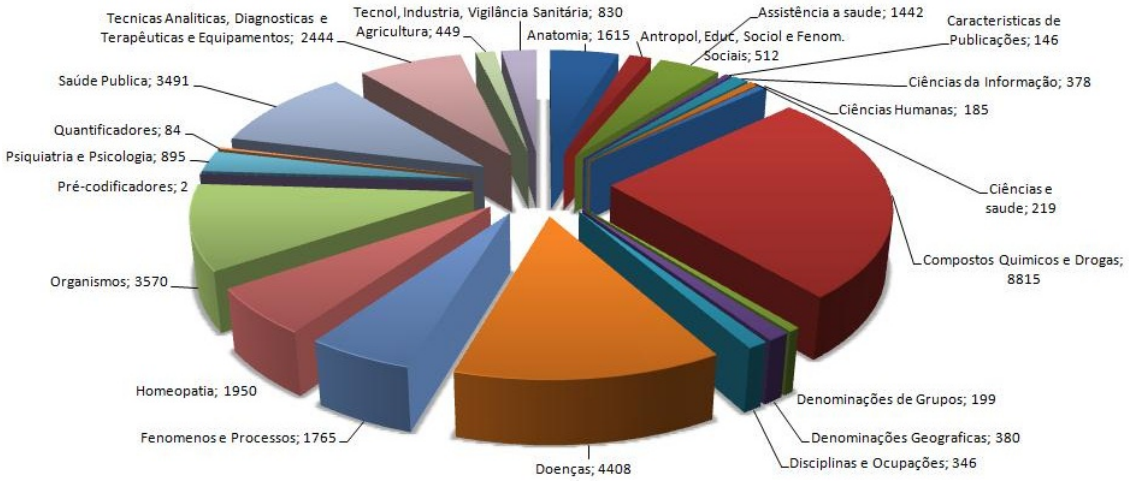

Figura 3.7: As Categorias que constituem o DeCS na versão 2010

O DeCS é um léxico trilingue (inglês, espanhol, português (Br)). As versões em espanhol e português do DeCS são exportadas para o Metathesaurus, e distribuídas como MeSH Espanhol e MeSH Português (Br)

3.1.2.2 Extracção de Informação

Como vimos anteriormente o DeCS está organizado. Não é uma colectânea criada por voluntários, mas sim um trabalho de profissionais. Por isso, os problemas encontrados na extracção dos termos do Wikipédia não são encontrados aqui. Não existe a possibilidade do *crawler* perder-se na complexidade do grafo direccionado, uma vez que todo o léxico foi criado e estruturado a pensar no fácil acesso à informação Médica.

[12] www.Decs.bvs.br

Também, o problema de variações na estrutura das paginas entre termos não é tanto frequente. Tudo isto facilita muito o trabalho de extracção da informação contida no DeCS. Apenas é preciso ter em consideração que o DeCS é um léxico poli-hierárquico. Devido à natureza multidisciplinar, um conceito pode estar contido em mais que um ramo da hierarquia, como é possível ver na figura 3.8 onde o termo Homeopatia é acessível por dois caminhos.

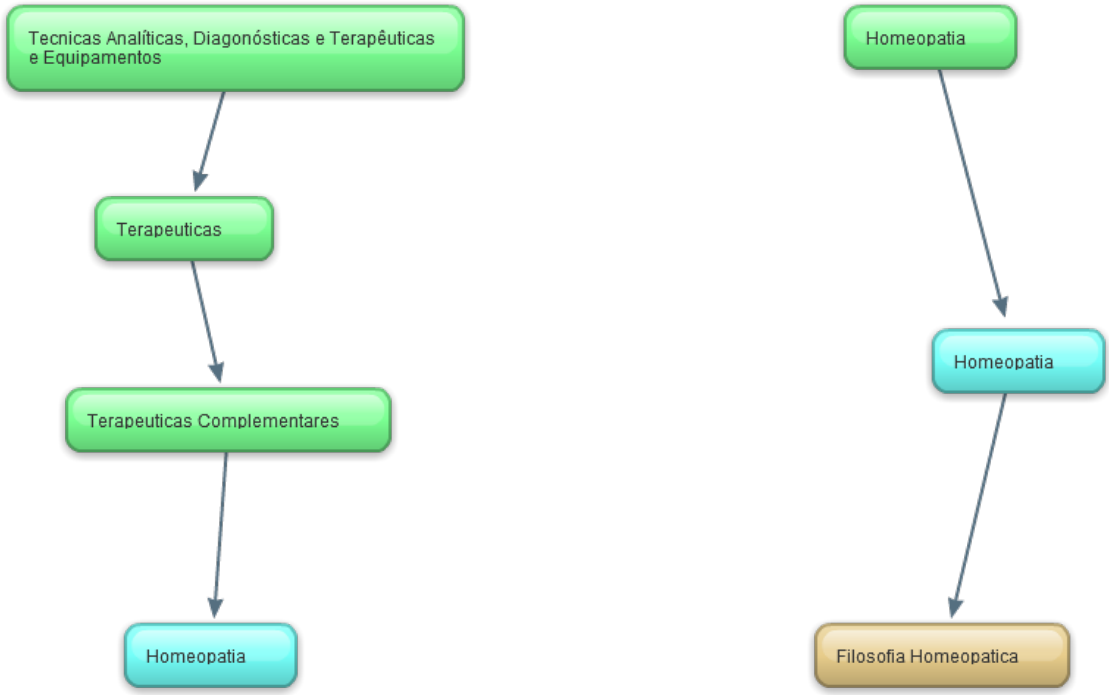

Figura 3.8: Diferentes ramos onde se insere o termo Homeopatia

Após ponderar todos os aspectos da estrutura do DeCS o *crawler* foi adaptado para extrair os artigos do DeCS que contem muita informação útil importante de recolher, como por exemplo informação sobre sinónimos, palavras relacionadas, abreviaturas e tradução do termos nas línguas inglês, e espanhol, assim sendo o caminho desde a categoria inicial até ao termo.

Como nos dicionários anteriores, a informação é guardada num ficheiro XML, cujo DTD é igual ao mencionado anteriormente.

3.2 Unificação do Dicionário

Uma vez recolhida a informação de todas as seis fontes, em que apenas a extracção do Wikipédia, Wikcionário e DeCS são abordados no âmbito desta tese,é importante reportar que o léxico que está na base deste projecto é composto por mais três fontes de informação: o Priberam, o Médicos de Portugal, e o Glossário Medico.

O facto de estarmos a construir um léxico médico unificado implica que incoerências e incorrecções sejam detectadas e corrigidas. Pois um léxico médico como descrito nos capítulos anteriores tem que ser claro e rigoroso. Para isso, é importante eliminar quaisqueres ambiguidades.

Através das fontes usadas no projecto foram detectados dois problemas que necessitam análise. Os erros ortográficos são um facto uma vez que usamos fontes de informação colaborativas em que qualquer pessoa pode contribuir com o seu conhecimento. Segundo, o aparecimento de termos unicamente brasileiros, como por exemplo cisto (Br) comparado com quisto (Pt) é um problema. Ambas as palavras significam o mesmo no entanto mas com ortografia diferente.

A criação de um léxico actual implica também que este seja a par da evolução linguística e portanto é indispensável que sejam aplicadas regras morfológicas para que o léxico que seja compatível com o novo acordo ortográfico que visa encurtar a distância entre o português lusitano e o português brasileiro [23].

Todo o trabalho de análise correcção e unificação é um processo que requer muita atenção e cuidado. Pois pode pôr em causa a reputação de todo o projecto. É por isso que todo esta etapa é feita manualmente, para garantir o rigor geral de todo o léxico. Esta parte é o trabalho da estudante de Doutoramento Isabel Marcelino.

Uma vez analisadas todas as bases de conhecimento obtemos um léxico com cerca de 55 000 termos distribuídos pelas diferentes fontes analisadas como podemos ver na figura 3.9, e mais 30 000 definidos como termos relacionados, sinónimos e outras relações entre termos. O que faz com que na base de todo o projecto está um léxico unificado com cerca de 85 000 termos.

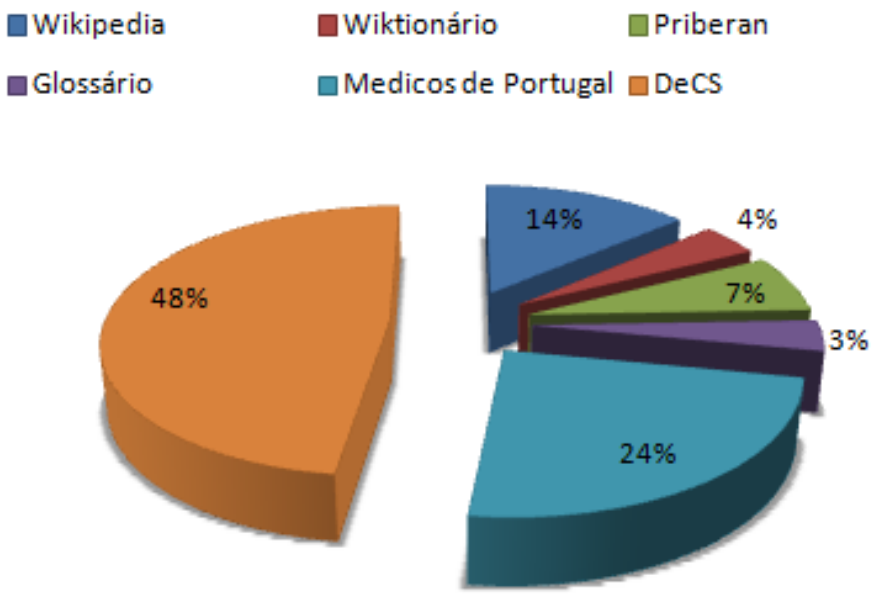

Figura 3.9: Composição do léxico unificado

Capítulo 4

Desenvolvimento da Interface

O objectivo deste trabalho é disponibilizar ao publico um léxico médico unificado rigoroso e fiável. Para isso é necessário criar plataformas que possibilitam a qualquer utilizador o acesso à informação, forma simples e intuitiva. De facto a informação só tem interesse para o utilizador se for acessível e perceptível, chamando a atenção do utilizador para o que realmente o interessa.

Criar uma interface capaz de agradar ao utilizador é mais do que posicionar botões ou criar menus. Uma interface é em primeiro lugar uma ponte de ligação entre o utilizador e a aplicação, ou seja o *design* de uma interface não só é sobre como se apresenta a informação mas também como ela funciona, não é só escolher cores ou botões mas também escolher as ferramentas certas para o trabalho [24].

A interface é um aspecto muito importante de uma aplicação, e em especial aplicações como o UMLP porque estão abertas ao público em geral. O utilizador vê e interage com a interface, não com a aplicação que está por detrás de todo o projecto [25]. Ter este elemento da aplicação certo vai ter um grande impacto na maneira como os utilizadores vão gostar de usar o produto.

O léxico está disponível em duas plataformas distintas, uma versão online, mais detalhada e com mais recursos, e uma versão para dispositivos móveis para que o utilizador possa aceder a informação em qualquer lugar. particular, os profissionais da saúde têm uma necessidade de movimentação importante no seio de um hospital, por exemplo.

4.1 Implementação para PC

Antes de pensar na nossa interface fomos analisar vários projectos, de forma a garantir a melhor qualidade da nossa interface

4.1.1 Aplicações Existentes

4.1.1.1 Médicos de Portugal

O projecto Médicos de Portugal tem como objectivo melhorar o estado da saúde em Portugal, disponibilizado e dinamizando um canal de comunicação na internet sobre saúde, disponível para todos os portugueses desde utentes a profissionais da saúde e solidariedade [1].

O Médicos de Portugal tem uma variedade de serviços disponíveis gratuitamente, desde informativos: Canal de Utentes; Canal de Médicos; Canal de Solidariedade; Pesquisas de médicos e instituições; NewsLetter; Glossário; Farmácias, e serviços de classificados: Empregos; Equipamento e materiais; Voluntariado.

Para os médicos também oferece uma área para a publicação de artigos científicos, assim como a possibilidade de registar instituições. Para este projecto analisar a área do glossário é importante. O Médicos de Portugal é um vasto projecto, e por isso encontrar o glossário não é propriamente uma tarefa intuitiva. Mas uma vez descoberta temos uma pagina como aquela apresentada na figura 4.1. A interface é simples e prática, a pesquisa pela ordem alfabética é visível no início da página, e a procura por um termo específico está logo abaixo, sendo fácil para qualquer pessoa procurar por um termo.

Uma vez efectuada uma pesquisa, por exemplo "Anemia" os resultados são imediatamente apresentados. A pesquisa é rápida. No entanto, o excesso de resultados é notório. São apresentados 89 termos que variam desde "Cancro Gástrico" a "Tricocefalose" para o termo "Anemia".

Como podemos ver na figura 4.2, a informação referente à categoria gramatical, tradução do termo nas línguas inglesa e francesa, e o adjectivo relacionado (anémico) são disponibilizadas.

[1] http://medicosdeportugal.saude.sapo.pt

Figura 4.1: Página inicial do Glossário

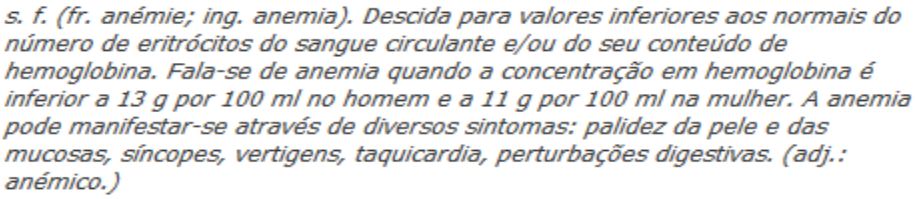

Figura 4.2: Definição de Anemia

4.1.1.2 Dicionário Priberam da Língua Portuguesa - DPLP

O Dicionário Priberam da Língua Portuguesa é um dicionário de português europeu. É um dicionário geral, não estando limitado ao domínio da Medicina, não incluíndo termos na sua variante brasileira. No entanto, já possui uma versão que permite consultar e comparar a grafia das palavras antes e depois da aplicação das regras do novo acordo ortográfico [2].

É relativamente simples de usar e qualquer pessoa está familiarizada com a sua estrutura, pois é semelhante a muitas páginas Web de pesquisa, como podemos ver na figura 4.3

O DPLP permite pesquisar por termo ou na definição de cada termos. A caixa de pesquisa possui a propriedade de auto-completar o termo que está a ser escrito para ajudar o utilizador. É um pormenor muito útil especialmente as palavras complexas.

[2] http://www.priberam.pt/

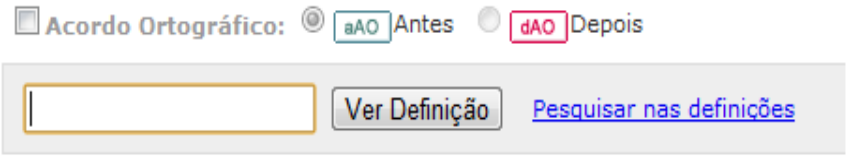

Figura 4.3: Caixa de pesquisa

Como podemos ver na figura 4.4

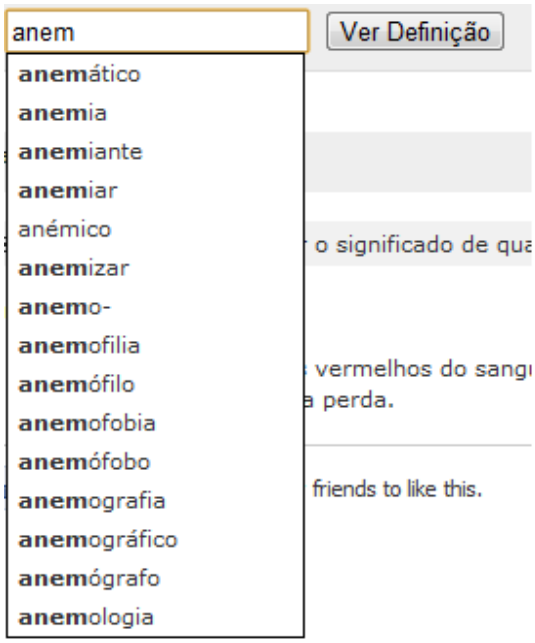

Figura 4.4: Exemplo de sugestões para completar o termo

Se o acordo ortográfico estiver activo, então à medida que a palavra é introduzida na caixa de texto, vão aparecendo as duas formas, antes e depois do acordo ortográfico, como podemos ver na figura 4.5.

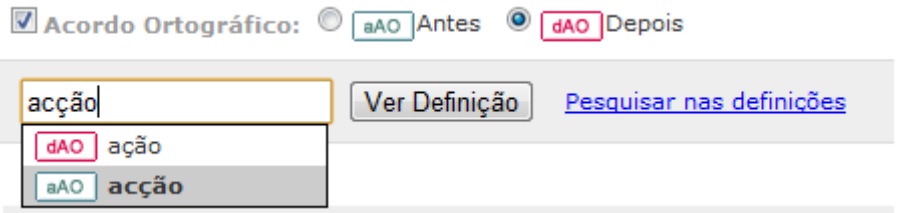

Figura 4.5: Antes e depois do acordo ortográfico

Uma vez feita a pesquisa são apresentados os resultados. O DPLP procura pela palavra exacta, e não por uma aproximação como no caso do Médicos de Portugal,

o que leva a que neste caso não seja apresentada uma lista de termos mas sim directamente a definição quando o termo é encontrado. A estrutura básica do DPLP inclui informação sobre a grafia, pronúncia, origem da palavra, classificação gramatical, definição, sinónimos e antónimos (identificados respectivamente por $=$ e $\neq$), número, género, superlativos e variantes. No caso de "Anemia" como no exemplo anterior o DPLP (como mostra a figura 4.6), também apresenta um gráfico referente ao número de vezes que a palavra foi consultada. Termos relacionados não são apresentados na mesma página, são apresentados numa lista de palavras cada uma com a definição gramatical e também um pouco da sua definição, para que o utilizador posso saber do que se trata antes de ver com detalhe um termo, podendo assim ir directamente à definição que lhe interessa sem perder tempo analisando as palavras relacionadas que não lhe despertam interesse (ver figura 4.6).

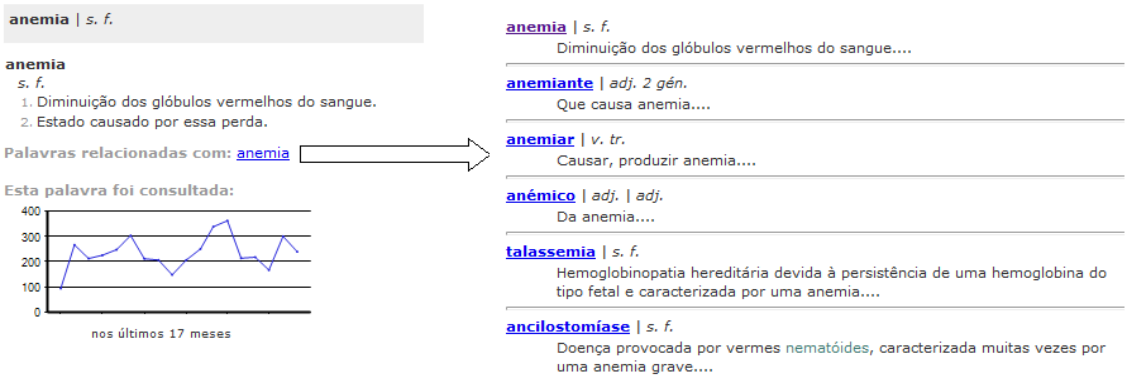

Figura 4.6: Extracto da definição apresentada pelo DPLP

Nas pesquisas que efectua um utilizador, este deve ter presente que a nomenclatura do DPLP, assim como a de qualquer outro dicionário, não é exaustiva. Como o DPLP é um dicionário electrónico em constante actualização e aperfeiçoamento, é normal que uma palavra pesquisada não pertença ao domínio do DPLP, De facto, o DPLP através do FLiP (Ferramentas para a Língua Portuguesa),uma ferramenta que disponibiliza diversos produtos e serviços na área do processamento da língua natural, consegue sugerir outras formas gráficas que se aproximam da forma digitada, possibilitando assim alguma correcção de erros por parte do utilizador.

A pesquisa na definição é uma funcionalidade bastante útil se o utilizador está à procura do contexto em que uma palavra ou um conjunto de palavras é utilizador.

4.1.1.3 Wikipédia

O Wikipédia quase despensa apresentações, embora não seja um dicionário, mas sim um enciclopédia online. No entanto é um dos sites mais procurados quando se pretende encontrar uma definição.

O sistema de procura assim como o DPLP também possui a propriedade de aparecer com uma lista de possíveis palavras à medida que o utilizador vai digitando o termo que deseja. Uma vez a pesquisa feita é apresentada caso exista, toda a informação para o termo, desde definições, características históricas, imagens e algumas referências externas. O conteúdo está também interligado através de hiperligações para palavras existentes no domínio do Wikipédia. É possível quando disponível, também navegar para o termo noutras línguas (inglês, francês, etc.) [26]

4.1.2 Solução de Interface Apresentada

O "Escul@pio" é o nome dado a esta aplicação, encarregue de levar até ao utilizador o léxico unificado, e a possibilidade de incluir também o seu conhecimento ao léxico, através de vídeos, imagens, comentários, ou até mesmo alterações na definição de um termo.

Recorrendo às mais recentes tecnologias na criação de páginas para a Web, foi desenvolvido uma interface que oferece ao utilizador uma fácil, intuitiva e agradável experiência de utilização, não despejando informação no monitor, mas sim organizando-a de maneira que o utilizador preste a sua atenção no que lhe é realmente importante. Pensamos assim ter criado um site Web apelativo e de fácil utilização sem descuidar as suas funcionalidades.

A Figura 4.7 mostra a página inicial da aplicação. É logo possível verificar que ao contrário dos dicionários descritos no capítulo anterior, o Escul@pio não enche a página com informação desnecessária. É uma página simples com apenas alguns botões que o utilizador entende facilmente. Assim, o utilizador pode rapidamente fazer uma pesquisa sem ser necessário perder tempo a analisar toda a informação que para ele é desnecessária.

O Escul@pio possui algumas opções na pesquisa por termo. A figura 4.8 mostra as diferentes maneiras de pesquisar um termo. A pesquisa pode ser feita pelo termo exacto, ou por uma palavra que contenha esse termo. Por exemplo, se a opção de "Por termo

Figura 4.7: Página principal do Escul@pio

exacto" não estiver activa (ver figura 4.8), na pesquisa por "Anemia" a palavra "Anemia Aplástica" também será inserida no grupo de resultados, uma vez que o termo "Anemia" faz parte da sua formação. A Opção de "Na definição", irá procurar pela utilização do termo nas definições, apresentado assim todos os termos que lhe façam referência. Assim, como vimos nas aplicações da Priberam e do Wikipédia, também o Escul@pio tem o sistema de autocompletar o termo à medida que este vai sendo digitado. Isto é uma grande ajuda para os termos complexos que o domínio da Medicina é abundante (ver figura 4.9).

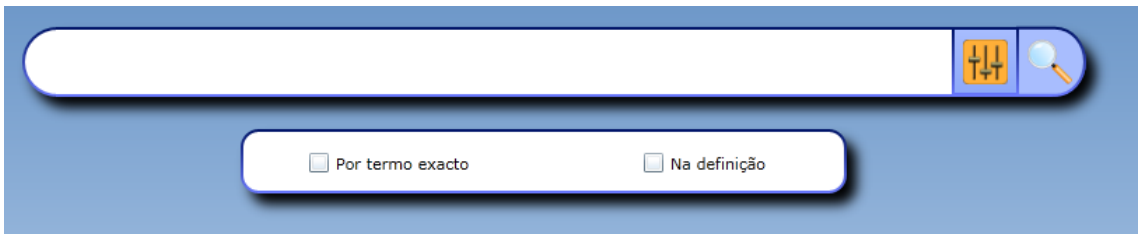

Figura 4.8: Caixa e filtros de pesquisa

O léxico usado por detrás desta interface para muitos termos possui a sua tradução nas variantes inglês, francês e espanhol. Quando disponível, é natural que a interface permita ao utilizador que procure por um termo noutra língua. Por exemplo, quando um utilizador apenas sabe o nome de uma doença em inglês (por exemplo *Hematology*),

Figura 4.9: Exemplo do sistema autocompletar os termos

entao, escrevendo na caixa de pesquisa o termo mais a designação da lingua (neste caso "[eng]"),o Escul@pio entende que o termo é em inglês e vai procurar os termos que existam em todo o léxico cuja sua tradução em inglês seja *Hematology*. Esta função é muito útil por exemplo para estudantes estrangeiros que tenham dificuldades na escrita do português.

O Escul@pio possui um léxico relativamente grande com cerca de 85 000 termos. Encontrar e produzir resultados consoante a pesquisa efectuada não é uma tarefa difícil. O difícil é tirar partido dos resultados obtidos. Por isso o modo de visualização é muito importante, pois grandes quantidades de informação podem tornar-se confusos e pouco perceptivas. A utilização de uma visualização em três dimensões pode facilitar estes problemas. Pois, acrescentado mais uma dimensão à representação de dados, tornado assim possível uma utilização mais eficiente do espaço limitado do monitor, além de que é mais atractivo ao utilizador, como podemos ver na figura 4.10.

Uma vez efectuada a pesquisa, vai ser criada o grupo de resultados. Aqui é que o Escul@pio começa a fazer a grande diferença com os outros dicionários online. No que respeita ao modo como é apresentada a informação, a interface tenta sempre ter um visual agradável, assim como funcional. Na pesquisa pelo termo "Anemia" são apresentados um total de 63 termos naqual a palavra "Anemia" faz parte. Logo são muitos termos para o utilizador ver de uma só vez, e apresentar uma lista de palavras

não é muito agradável nem prático porque o utilizador pode não conseguir encontrar a palavra que deseja sabendo que a sua atenção pode não estar focada num ponto mas sim numa lista de palavras. Para resolver este problema o Escul@pio apresenta dois modos de visualização para o grupo de resultados. O primeiro apresentado na figura 4.10, mostra os resultados numa estrutura em 3D, que usa parâmetros visuais para facilitar a compreensão do utilizador, usando métodos de focagem para trazer para o centro da atenção do utilizador apenas uma parte da informação disponível, não despejando tudo de uma vez. Isso faz com que o utilizador apenas prenda a sua atenção no termo que se encontra no meio da pagina. À medida que vai navegando, os termos vão mudando de posição de maneira a que o utilizador possa procurar em todos os valores apresentados o que lhe interessa. Em particular, os resultados são separados em grupos de dez elementos para facilitar a navegação.

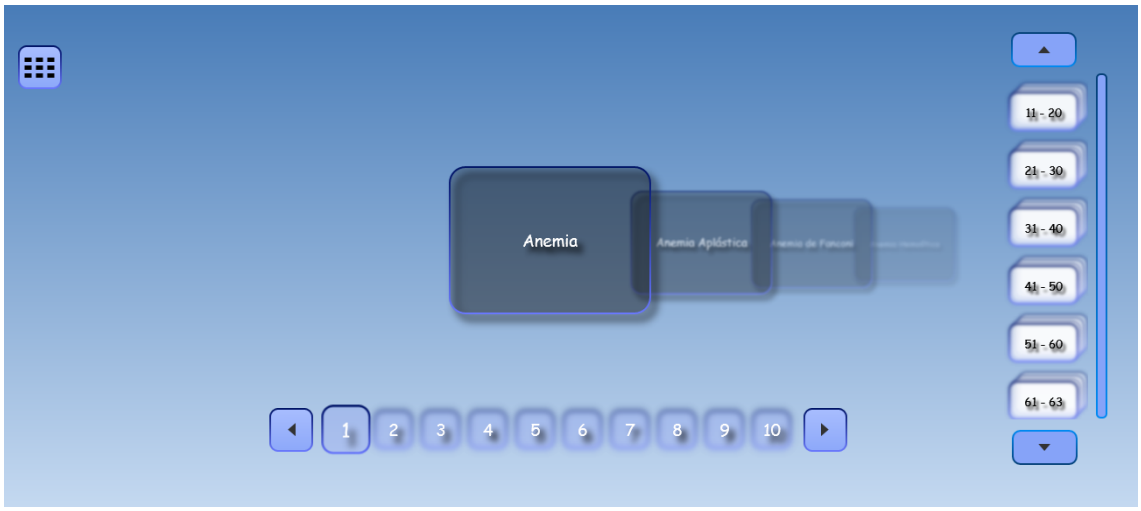

Figura 4.10: Grupo de resultados. Visualização focando apenas um elemento

O botão no canto superior esquerdo serve para mudar para outra forma de visualização dos resultados, uma vez que esta visualização pode levar a que o utilizador perca mais tempo na procura do termo que lhe interesse. Um outro modo de visualizar os dados foi concebida como podemos ver na figura 4.11, em que os dados são apresentados numa tabela de valores, sem haver necessidade de carregar apenas dez resultados, mostrando assim todas as entradas do grupo de resultados.

Esta visualização não usufrui das propriedades de uma vista em 3D, e foi feita a pensar nos utilizadores que preferem visualizações simples, práticas e que lhes mostre toda a informação. No entanto, a interface não mostra uma lista de palavras.

Figura 4.11: Grupo de resultados. Visualização em colunas

Pois isso poderia não ser muito productivo, podendo levar a que o utilizador não consiga encontrar o termo que pretende na imensidão de resultados. Os resultados são apresentados em colunas, espaçados entre si, que reagem ao posicionamento do rato para que o utilizador consiga acompanhar com os resultados obtidos, mantendo assim um visual agradável ede facíl compreensão.

Uma vez encontrado o termo pretendido, o utilizador pode ver o seu conteúdo unificado, i.e. as várias definições, informação gramatical, etimologia, sinónimos, palavras relacionadas, abreviações e símbolos, tradução do termo nas línguas inglesa, francesa e espanhola. Como podemos ver na figura 4.12, à esquerda temos as definições apresentadas, e à direita é apresentada toda a informação complementar do termo.

Esta parte da interface é muito importante, pois é aqui que a informação do léxico referente ao termo pesquisado é apresentada. Um léxico unificado significa que um termo pode ter mais do que uma definição, ou um termo pode ser definido através de relações de palavras ou ser proveniente de várias fontes. O problema baseia-se no facto de mostrar ao utilizador tanta informação sem que esta fique confusa.

Do lado esquerdo são apresentadas as diferentes definições do termo. As definições são agrupadas pelas suas origens. Por exemplo as definições do Wikcionário estão todas dentro do mesmo conjunto, e as definições encontradas no DeCS são postas noutro conjunto. Assim, o utilizador tem rapidamente noção de onde veio a definição que está a ver. As palavras que definem o termo através de uma relação semântica ou palavras relacionadas, também aparecem no conjunto de definições. No entanto, o

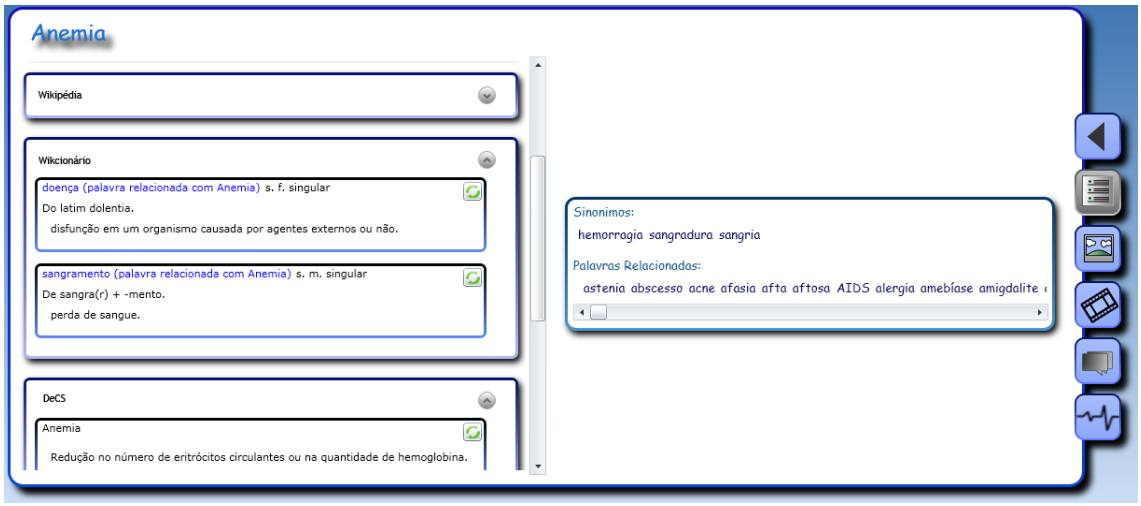

Figura 4.12: Disposição da informação referente ao termo unificado

termo aparece em cor diferente (azul para palavras relacionadas e verde para sinónimos), assim como uma pequena informação para que o utilizador não seja conduzido em erro, nem confunda o significado do termo. No exemplo de "Anemia" como podemos ver na figura 4.12, o léxico apresenta as definições que encontra para o termo, neste caso, sete termos no total que definem o termo "Anemia", na figura apenas estão visiveis três do total das definições, divididos por seis fontes de vocabulário, o Wikcionário define o termo através das palavras relacionadas "Doença" e "Sangramento".

A informação técnica sobre o termo e a sua fonte também não foram deixadas de fora. Cada definição, em cada termo unificado, possui a sua própria "ficha técnica" com informação da sua origem, data do documento original, e data da última actualização. No entanto, esta informação, embora muito importante, não é o que o utilizador procura inicialmente. Por isso não é logo visível na página. Encontra-se "escondida por detrás" da definição de cada termo. Assim caso o utilizador esteja interessado em ver quando foi feita a última actualização ou até mesmo visitar a página original do termo, pode fazê-lo.

Do lado direito são apresentados os dados referentes a sinónimos, antónimos, palavras relacionadas, traduções (inglês, francês e espanhol), abreviaturas e símbolos, sempre que estes estejam disponíveis. Inicialmente é apresentado o resultado da unificação, i.e. toda a informação de todos os termos que compõem a unificação. Se o utilizador estiver interessado em apenas uma definição, basta clickar em cima da definição e imediatamente os dados são actualizados para apenas o que diz respeito

à definiçãoem questão.

As palavras que compõem esta informação de relação entre termos podem conter informação adicional, por exemplo, categoria gramatical, género e número, tipo de utilização (por exemplo termos de utilização popular). Para ver esta informação basta pousar o rato em cima da palavra e caso haja informação ela é apresentada. Caso o utilizador fique interessado em alguma definição destes termos, caso exista no léxico, com um click em cima do próprio termo é apresentada outra estrutura de informação, com os dados do termo correspondente.

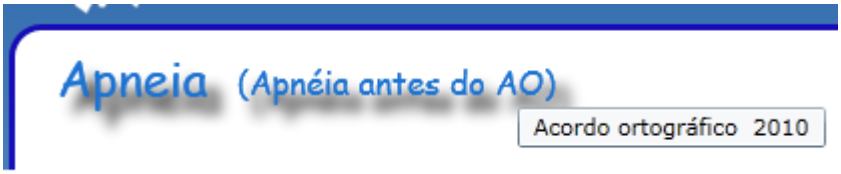

Figura 4.13: Exemplo de pesquisa por termos alterado pelo acordo ortográfico

Também o acordo ortográfico não foi esquecid. No entanto, a maioria dos portugueses não está habituado às novas regras gramaticais. É portanto esperar-se que o utilizador use as duas formas gramaticais da palavra. Por exemplo, quando um utilizador pesquisar por "Apnéia" o dicionário vai mostrar-lhe o termo correcto "Apneia" mas vai também deixar a indicação que a palavra foi alterada com o acordo ortográfico como podemos ver na figura 4.13

Um aspecto muito importante e quase indispensável num dicionário electrónico, é a inclusão de informação multimédia, imagens e vídeos. Pois, ajudam a compreender melhor o significado do termo. O Escul@pio não deixa esta parte de fora e possui uma secção de imagens e outra de vídeos com as respectivas legendas. Como podemos ver na figura 4.14, à direita da figura está a parte dos vídeos e à esquerda as imagens. Os termos no léxico são definidos por várias fontes, várias definições o que leva a que um termo possa não ter qualquer imagem ou vídeo, ou ter vários de cada.

Na figura 4.14 à esquerda podemos ver como os vários elementos são organizados num estrutura em 3D mantendo assim apenas um elemento em foco, enquanto os outros elementos encontram-se mais distante. Para cada elemento multimédia existe uma legenda que se encontra por baixo. A informação técnica também está presente, por detrás do elemento, de onde veio e quando foi adicionado, com uma hiperligação caso o elemento seja originário de uma pagina Web. Para as imagens existe também a possibilidade de fazer um *zoom*, para ver melhor algum detalhe. usando o *scroll* do

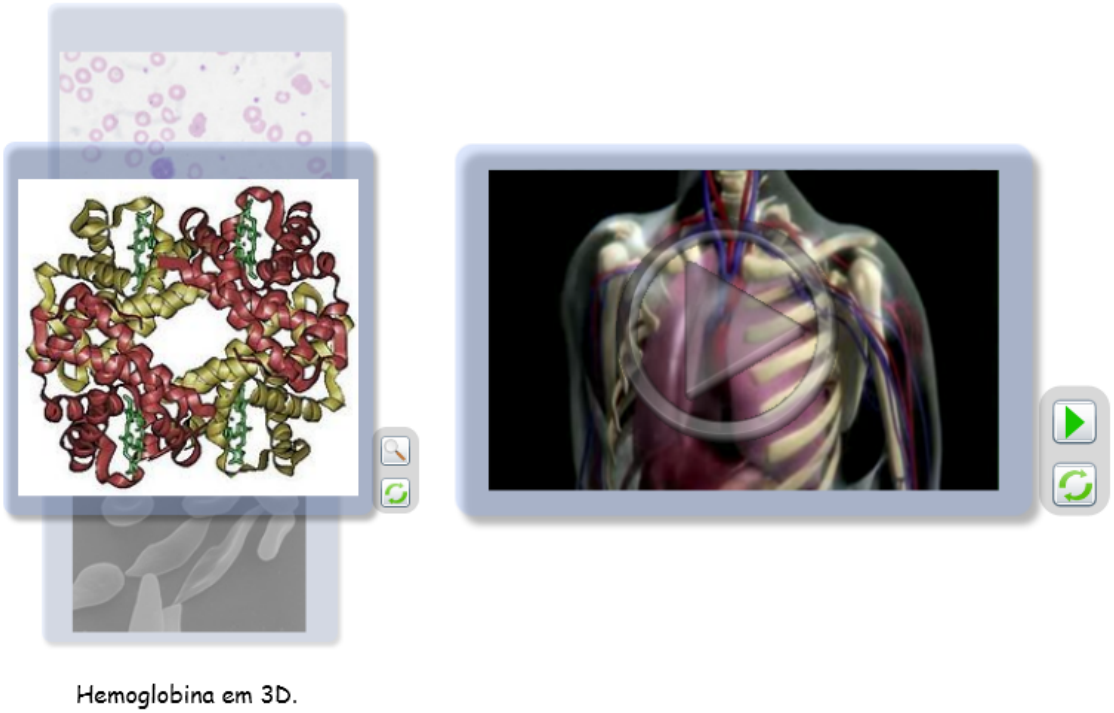

Figura 4.14: Elementos multimédia para o termo Anemia

rato é possível aumentar ou diminuir a imagem, ou arrastá-la para de um lado para o outro.

Adicionar novos elementos é um aspecto que não pode faltar quando falamos de um dicionário colaborativo. O utilizador pode fazer o upload do material que possui, Este é registado como originário do utilizador, para que toda a gente saiba de onde veio o elemento.

A cada termo um utilizador pode deixar o seu comentário. A opinião do utilizador é sempre importante e útil. É através do seu comentário que se pode completar alguma informação que não tenha espaço noutra secção do termo, ou que os outros utilizadores podem ver outro ângulo. Este espaço serve como ponto de conversa entre os utilizadores. Aqui podem-se discutir pontos de vista ou até mesmo discutir a sua experiencia pessoal, tornado assim o léxico mais pessoal.

Na figura 4.15 podemos ver um exemplo de comentários ao termo Anemia. Os comentários estão disponíveis para qualquer utilizador, no entanto apenas os utilizadores registados podem deixar o seu comentário. Cada comentário tem a informação do seu autor, nome e foto, assim como a data a que foi criado, e os valores da votação sobre os gostos dos utilizadores. Cada utilizador pode votar uma vez para cada termo. No

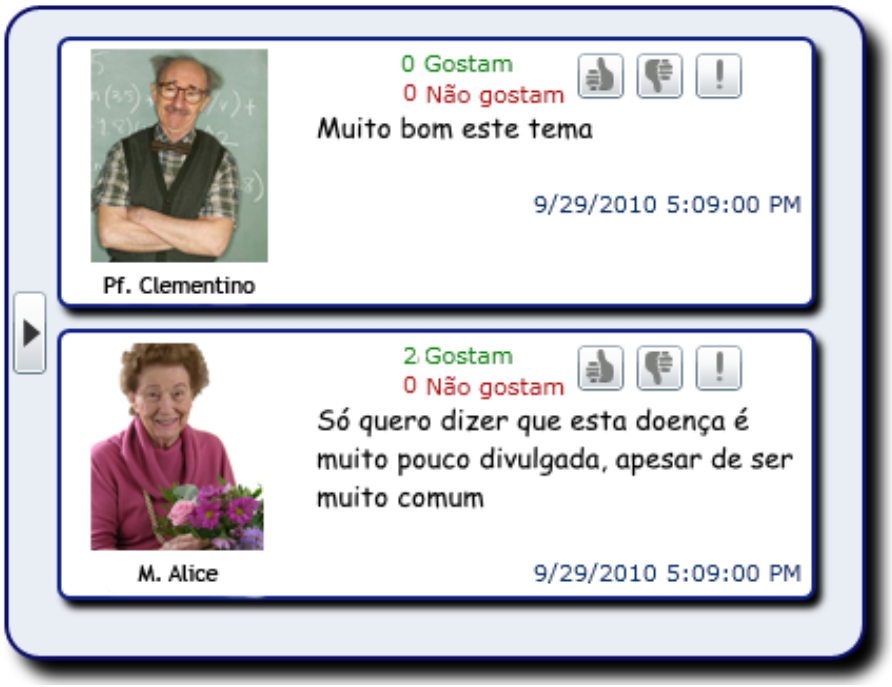

Figura 4.15: Exemplo de comentários a um termos

entanto, a qualquer altura pode mudar o valor do seu voto.

O Escul@pio também tem uma secção com informação sobre popularidade do termo, em que é apresentadq informação sobre quantos utilizadores procuraram a definição de um termo nos últimos tempos. Assim, um utilizador pode saber se um termo é muito visitado, e qual foi a altura em que mais utilizadores pesquisaram esse assunto, e assim tirar conclusões sobre a actualização da informação presente.

Um aspecto muito importante num site colaborativo é o login de utilizadores. Para que seja possível a um utilizador deixar o seu comentário, adicionar uma imagem ou vídeo, entre outras funções precisa estar registado e entrar com a sua conta de utilizado.

O registo é muito simples e fácil sendo preciso inserir o nome e apelido, um e-mail válido e uma palavra-chave, para poder efectuar o login. Os campos podem mais tarde ser alterados na edição de perfil, e também adicionar mais informação pessoal, como data de nascimento, sexo, país, cidade onde vive e uma foto. Uma vez feita o login as opções exclusivas a utilizadores registados ficam automaticamente disponíveis e assim o utilizador pode tirar o máximo partido de toda a interface.

Na figura 4.16 é visível as janelas de login e de edição de perfil, um layout bastante simples e muito prático, sendo fácil e rápido, não tornando-se aborrecido ao utilizador

Figura 4.16: Janelas de login e de edicção de perfil

actualizar os seus dados.

4.2 Implementação para Dispositivo Móvel

Cada vez mais os telemóveis e dispositivos móveis estão a ganhar terreno no acesso à internet e por conseguinte à informação nela presente. Isto tem vindo a apresentar novos desafios na construção de interfaces capazes de responder às exigências do mercado.

Trabalhar para dispositivos móveis significa ter em atenção requerimentos especiais, que em PC por norma não são problemas, e o mais provável é nunca serem levados em conta. O reduzido tamanho do monitor, a memória e o poder de processamento impõem limites ao que pode ser alcançado nos dispositivos móveis. Portanto criar uma interface para este tipo de dispositivos requer muita engenhosidade para trabalhar com limitações impostas e mesmo assim conseguir corresponder às exigências do utilizador, que procura uma aplicação que seja simples, intuitiva e fiável.

4.2.1 Exemplos de Aplicações Móveis

4.2.1.1 DPLP

O DPLP já visto anteriormente, também possui uma aplicação para iPhone e o iPod touch, que fornece consultas ao dicionário por meio de uma ligação à internet. A consulta pode ser feita com ou sem as alterações previstas no acordo ortográfico de 1990. A aplicação permite consultar sinónimos e antónimos, e em alguns casos é também possível consultar informação sobre a origem da palavra e a sua pronúncia. Esta aplicação permite ainda ver a palavra do dia e a mais pesquisada. Possibilita ainda a criação de uma lista de palavras favoritas que são guardadas no dispositivo e que podem ser consultadas mesmo na ausência de uma ligação à internet [27].

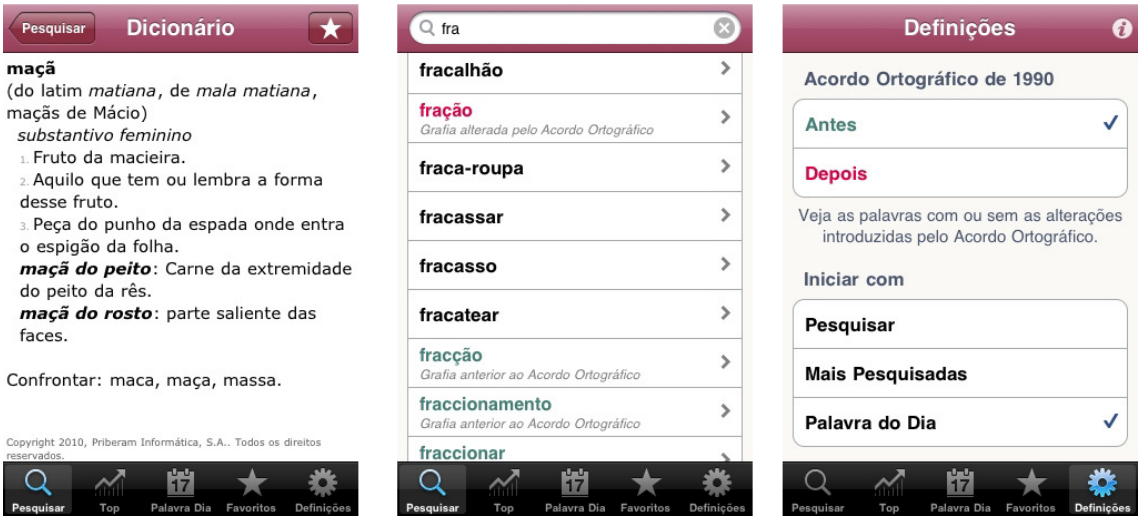

Figura 4.17: *Screenshots* da aplicação do DPLP para o iPhone

A figura 4.17 mostra a aplicação em três *screenshots* que mostram o funcionamento da aplicação no dispositivo móvel e também como a informação é apresentada.

4.2.1.2 Clustering e Sumariando Documentos Médicos

Uma aplicação médica no sentido de dispositivos móveis prende-se com o facto de facilitar o acesso informação sobre testes clínicos, estudos e literatura científicos no geral, e também de facilitar a mobilidade dos profissionais de saúde nos hospitais. Devido às limitações dos dispositivos é necessário seleccionar a informação crucial e apresentá-la sintetizada.

O artigo descrito em [28] tem como proposta usar um sistema de sumarização de vários documentos e agrupá-los para a recuperação de informação para dispositivos móveis. O resultado final é um sistema que oferece um sumário de todos os clusters e mostra semelhanças entre documentos. A figura 4.16 mostra um exemplo de como os resultados são expostos. Podemos ver a representação de cada cluster, com o seu nome e o número de documentos que eles contêm, e um pequeno sumário baseado na semelhança. O utilizador pode seleccionar um cluster ou fazer uma nova pesquisa. Quando um cluster é seleccionado, é mostrado o título dos vários documentos e o seu sumario como podemos ver na figura da direita.

Figura 4.18: *Screenshots* que mostram os resultados usando um protótipo. A imagem da esquerda mostra os clusters e a imagem da direita o conteúdo de um dos clusters

4.2.2 Solução de interface apresentada

Os benefícios de uma aplicação móvel deste tipo já estão descritos em capítulos anteriores deste trabalho, e por isso houve sempre uma enorme vontade de fazer uma interface para dispositivos móveis capaz de levar até ao utilizador um serviço prático e fácil de usar.

A versão móvel do Escul@pio apresenta uma interface muito semelhante à da versão PC como podemos ver na figura 4.19. Pois assim o utilizador já está familiarizado com os funcionamento da aplicação. No entanto devido às limitações físicas dos aparelhos móveis, a aplicação não suporta todas as funcionalidades.É um projecto ainda em

desenvolvimento, com um objectivo futuro muito ambicioso, com o alvo de se tornar uma ferramenta indispensável na vida de um profissional de saúde [29].

A primeira versão do Escul@pio para sistemas móveis permite ao utilizador pesquisar no léxico unificado as várias definições das diversas fontes presentes na unificação. Na figura 4.19 podemos ver um exemplo da aplicação a funcionar para a pesquisa do termo "Medicina". Uma vez feita a pesquisa, o léxico devolve um grupo de resultados cujo termo de pesquisa se aproxime graficamente dos termos encontrados. São então apresentados um grupo de resultados, uma espécie de cluster de definições para o termo como podemos ver na imagem ao centro, onde dentro de cada cluster está a definição unificada das várias fontes que constituem o léxico. A informação gramatical, etimológica e a definição são apresentadas logo. Pois, inicialmente, é isto que um utilizador procura numa pesquisa como podemos ver na imagem da direita, onde estão as várias definições do termo "Medicina". Depois, consoante a definição seleccionada existem algumas funcionalidades que o utilizador pode usar caso pretenda i.e. ver sinónimos, antónimos, palavras relacionadas ou traduções nas línguas inglês, francês e espanhol ou mesmo visualizar imagens ou vídeos , como podemos ver na figura 4.20, a esquerda temos uma demonstração da lista de palavras relacionadas com medicina, e a direita está um exemplo de uma imagem que pertence ao termo "Medicina".

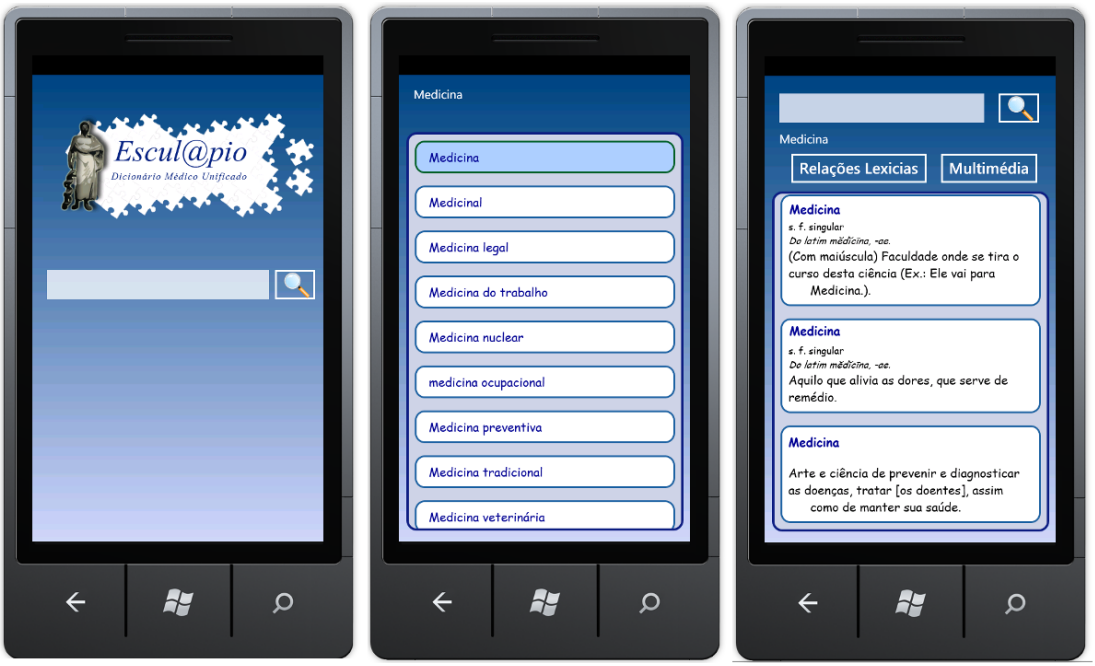

Figura 4.19: *Screenshots* da aplicação de dispositivos móveis, a esquerda é o ecrã inicial, no meio o grupo de resultados da pesquisa, e a direita o resultado da unificação do termo.

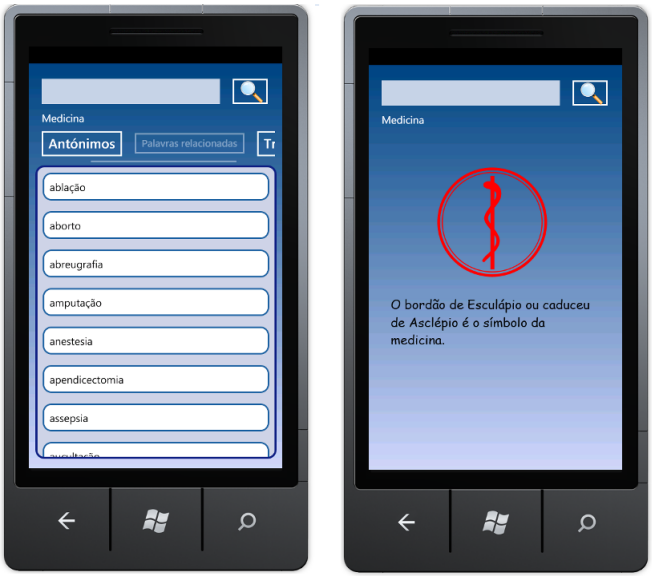

Figura 4.20: À esquerda as palavras relacionadas do termo, à direita uma imagem e respectiva legenda.

Capítulo 5

Conclusão e Trabalhos Futuros

5.1 Conclusão

Ţ

A internet é cada vez mais um poço de informação, e encontrar e extrair o que realmente interessa consoante o contexto em que o utilizador se insere é um desafio cada vez mais pertinente. Fontes de informação como o Wikipédia mostram-se muito complexos e diversificados. No entanto, foi possível criar meios de procurar e extrair informação relacionada com o domínio da saúde, de modo satisfatório, não dispensado a revisão linguística por parte de pessoas especializadas.

Criar duas plataformas semelhantes para dispositivos diferentes capazes de levar até ao utilizador o léxico de termos médicos unificado, foi desde sempre o alvo deste trabalho. Pois, é preciso criar meio de acessibilidade à informação para que esta se torne útil. Embora ainda sendo uma versão muito inicial, já é possível apresentar plataformas tanto para PC como para dispositivo móvel, capaz de levar até ao utilizador um léxico unificado, e apresentar os termos e as suas relações lexicais.

5.2 Trabalhos futuros

Ţ

Levar o Escul@pio mais além, é um objectivo. Actualmente, a plataforma é capaz de apresentar termos pesquisados e suas relações lexicais e gramaticais. No entanto, existem algumas ideias de futuras implementações muito interessantes para uma pla-

taforma deste tipo.

Criar um sistema de registo de utilizadores que seja fiável e dê garantias da seriedade de quem actualiza os conteúdos da base de dados, é algo a levar em conta. Usar um sistema capaz de ler o cartão do cidadão e criar o registo, pode ser uma maneira de resolver este problema.

Utilizar técnicas avançadas para analisar os tesauros existentes do DeCS e do Wikipedia, e a partir dai criar um novo tesauro (um Metathesaurus à semelhança do UMLS) mais correcto e capaz de responder melhor as exigências da plataforma, sendo este o próximo passo mais directo.

Integrar uma base de dados de medicamento fornecida pela INESC-ID(Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento), no projecto pode tambem ser uma maneira de o tornar muito mais completo e útil para qualquer tipo de utilizador.

Na Universidade de Évora está a nascer de um projecto do aluno Luís Borrego sobre a orientação do Prof. Doutor Paulo Quaresma, que visa criar ontologias em relatórios médicos. Também será uma boa ferramenta que juntamente com a nossa plataforma pode se tornar muito poderosa.

A ideia é levar este aplicação aos profissionais de saúde, ser algo que lhes vá facilitar a vida profissional. Portanto, criar uma área capaz de fornecer serviço de e-conteudos, onde seja possível partilhar documentos é uma boa ideia, para por exemplo, um professor poderia deixar artigos para os seus alunos, e também aceder a artigos no PubMed, levando a que um médico tenha a informação toda que necessita no seu dia-a-dia profissional, à distância de um botão. Uma secção para notícias especialmente da área é mais uma ideia muito útil.

Um outro desafio é a criação de um pequeno médico virtual, uma área que dada os sintomas seja capaz de, com base em sistemas de decisão e recorrendo às base de dados disponíveis, diagnosticar doenças frequentes.

Integrando todos estes componentes num serviço móvel vai permitir aos seus utilizadores uma rápido acesso a todo o tipo de informação, médica e relacionada com a saúde, criando assim meios para um serviço mais rápido e fiável por parte dos profissionais de saúde.

References

[1] Christof MÃijller Torsten Zesch and Iryna Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[2] C Lovis, R Baud, A.M Rassinoux, P.A Michel, and J.R Scherrer. Medical dictionaries for patient encoding systems: a methodology. *Artificial Intelligence in Medicine*, 14(1-2):201 – 214, 1998. Selected Papers from AIME '97.

[3] Chantelle Garritty and Khaled El Emam. Who's using pdas? estimates of pda use by health care providers: A systematic review of surveys. *J Med Internet Res*, 8(2):e7, May 2006.

[4] Olivier Bodenreider. The unified medical language system (umls): Integrating biomedical terminology, 2004.

[5] Humphreys B.L. Lindberg, D.A. and McCray A. The unified medicla language system. In *Methods Inf Med*, pages 281–291, 1993.

[6] Anand Smith, Barry Kumar and Steffen Schulze-Kremer. Revising the umls semantic network. 2004.

[7] McCray Browne and Srinivasan. The specialist lexicon. 2000.

[8] M. Zabel G.Weske-Heck, A. ZaiÎš. The german specialist lexicon. 2002.

[9] *Towards a unified medical lexicon for French*.

[10] Pierre Zweigenbaum, Robert Baud, Anita Burgun, Fiammetta Namer, ÃĽric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Jean-FranÃğois Forget, Magaly DouyÃĺre, and StÃĽfan Darmoni. Umlf: a unified medical lexicon for french. *International Journal of Medical Informatics*, 74(2-4):119 – 124, 2005. MIE 2003.

[11] Bruno Cartoni and Pierre Zweigenbaum. Semi-automated extension of a specialized medical lexicon for french. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA).

[12] Bireme Descritores em ciencia da saude. http://decs.bvs.br.

[13] Mariangela Spotti Lopes Fujita Vera Regina Casaru Boccato. Avaliacao da linguagem documentaria decs na area de fonoaudiologia na prespectiva do usuario: estudo de observacao de recuperacao da informacao com protocolo verbal. 2006.

[14] Oren Etzioni. The world-wide web: quagmire or gold mine? *Commun. ACM*, 39(11):65–68, 1996.

[15] Christof Müller and Iryna Gurevych. Using wikipedia and wiktionary in domain-specific information retrieval. In *CLEF'08: Proceedings of the 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access*, pages 219–226, Berlin, Heidelberg, 2009. Springer-Verlag.

[16] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, 2005.

[17] Fernanda B. ViÃĽgas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations. pages 575–582. ACM Press, 2004.

[18] Satoshi Sakai Hiroshi Nakagawa Yoji Kiyota, Noriyuki Tamura and Hidetaka Masuda. Automated subject induction from query keywords through wikipedia categories and subject headings. In Bente Maegaard Joseph Mariani Jan Odjik Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Language Resources*

and Evaluation (LREC'08), Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[19] Jakob Voss. Collaborative thesaurus tagging the wikipedia way. *CoRR*.

[20] Michael Strube and Simone P. Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. In *AAAI'06: proceedings of the 21st national conference on Artificial intelligence*, pages 1419–1424. AAAI Press, 2006.

[21] Kotaro Nakayama, Takahiro Hara, and Shojiro Nishio. Wikipedia mining for an association web thesaurus construction. In Boualem Benatallah, Fabio Casati, Dimitrios Georgakopoulos, Claudio Bartolini, Wasim Sadiq, and Claude Godart, editors, *Web Information Systems Engineering âĂŞ WISE 2007*, volume 4831 of *Lecture Notes in Computer Science*, chapter 27, pages 322–334. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[22] Roy T. Fielding, Day Software, and Richard N. Taylor. Principled design of the modern web architecture. *ACM Transactions on Internet Technology*, 2:115–150, 2002.

[23] *Lola Geraldes Xavier A lingua portuguesa em evolucao: os Acordos Ortograficos*.

[24] Dmitry Fadeyev. User interface design in morden web application.

[25] J Raskin. The humane interface. 2000.

[26] Wikipedia. http://www.wikipedia.org.

[27] Priberam. http://www.priberam.pt/.

[28] *Clustering and Summarizing Medical Documents to Improve Mobile Retrieval*, 2008.

[29] Inc. Sun Microsystems. Mobile information device profile white paper. 2000.

Anexo

```
<!ELEMENT wiktionary (entry+)>
<!ATTLIST entry id CDATA #REQUIRED>
<!ELEMENT entry (word,source,url,etymology?,domain?,paths,category?,nu
  image?,synonyms?,antonym?,related_adj?,related_nouns?,related_word?,a
<!ELEMENT word (#PCDATA)>
<!ATTLIST word id CDATA #IMPLIED>
<!ELEMENT source (#PCDATA)>
<!ELEMENT url (#PCDATA)>
<!ATTLIST url doc_date CDATA #IMPLIED>
<!ATTLIST url search_date CDATA #IMPLIED>
<!ATTLIST url type CDATA #IMPLIED>
<!ELEMENT etymology (#PCDATA)>
<!ELEMENT domain (word+)>
<!ELEMENT paths (path+)>
<!ELEMENT path (#PCDATA)>
<!ELEMENT category (#PCDATA)>
<!ELEMENT number (#PCDATA)>
<!ELEMENT gender (#PCDATA)>
<!ELEMENT definition (#PCDATA)>
<!ELEMENT image (legend*,url)>
<!ELEMENT legend (#PCDATA)>
<!ELEMENT synonyms (synonym+)>
<!ELEMENT synonym (word,category*,number*,gender*,usage*,abbreviation*)
<!ATTLIST synonym id CDATA #IMPLIED>
<!ELEMENT antonym (word+)>
```

```
<!ELEMENT related_adj (word,gender*,translation*)>
<!ELEMENT related_nouns (related_noun+)>
<!ELEMENT related_noun (word,gender*,number*,usage*)>
<!ATTLIST related_noun id CDATA #IMPLIED>
<!ELEMENT related_word (word+,usage*)>
<!ELEMENT abbreviations (abbreviation+)>
<!ELEMENT abbreviation (word+,usage*)>
<!ELEMENT usage (#PCDATA)>
<!ATTLIST translation lang (en|fr|sp) #REQUIRED>
<!ELEMENT translation (word+)>
```